ISSN: 2612-226X

Studi AISV 11

IL PARLATO IN AMBITO MEDICO: ANALISI LINGUISTICA, APPLICAZIONI TECNOLOGICHE E STRUMENTI CLINICI

SPOKEN LANGUAGE IN THE MEDICAL FIELD: LINGUISTIC ANALYSIS, TECHNOLOGICAL APPLICATIONS AND CLINICAL TOOLS

a cura di
Barbara Gili Fivela, Sonia d'Apolito,
Anna Chiara Pagliaro, Vincenzo Sallustio,
Maria Luisa Fiorella, Marco Siniscalchi

AISV
Associazione Italiana Scienze della Voce

IL PARLATO IN AMBITO MEDICO: ANALISI LINGUISTICA, APPLICAZIONI TECNOLOGICHE E STRUMENTI CLINICI

SPOKEN LANGUAGE IN THE MEDICAL FIELD: LINGUISTIC ANALYSIS, TECHNOLOGICAL APPLICATIONS AND CLINICAL TOOLS

a cura di
Barbara Gili Fivela, Sonia d'Apolito,
Anna Chiara Pagliaro, Vincenzo Sallustio,
Maria Luisa Fiorella, Marco Siniscalchi

Milano 2023

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazioni fonologiche, fino alle applicazioni tecnologiche.

I testi sono sottoposti a processi di revisione anonima fra pari che ne assicurano la conformità ai più alti livelli qualitativi del settore.

I volumi sono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

Curatore/Editor
Cinzia Avesani (CNR-ISTC)

Curatori Associati/Associate Editors
Franco Cutugno (Università di Napoli), Barbara Gili Fivela (Università di Lecce), Daniel Recasens (Università di Barcellona), Antonio Romano (Università di Torino), Mario Vayra (Università di Bologna).

Comitato Scientifico/Scientific Committee
Giuliano Bocci (Università di Siena), Silvia Calamai (Università di Siena), Mariapaola D'Imperio (Rutgers University), Giovanna Marotta (Università di Pisa), Stephan Schmid (Università di Zurigo), Carlo Semenza (Università di Padova), Alessandro Vietti (Libera Università di Bolzano), Claudio Zmarich (CNR-ISTC).

© 2023 AISV - Associazione Italiana Scienze della Voce
c/o LUSI Lab - Dip. di Scienze Fisiche
Complesso Universitario di Monte S. Angelo
via Cynthia snc
80135 Napoli
email: presidente@aisv.it
sito: www.aisv.it

Edizione realizzata da
Officinaventuno
Via F.lli Bazzaro, 18
20128 Milano - Italy
email: info@officinaventuno.com
sito: www.officinaventuno.com

ISBN edizione digitale: 978-88-97657-70-5
ISSN: 2612-226X

Sommario

BARBARA GILI FIVELA, SONIA D'APOLITO, ANNA CHIARA PAGLIARO, VINCENZO SALLUSTIO, MARIA LUISA FIORELLA, MARCO SINISCALCHI
Introduzione 7

ISABELLA LAURA SIMONE
Language impairment in neurodegenerative diseases 13

SERGE PINTO
Speech motor control from its impairment: the relevance of studying dysarthria in Parkinson's disease 29

PARTE I
Disturbi del parlato e della fluenza

DILETTA SCIANCALEPORE, VALENTINA LAVERMICOCCA, TERESANNA ILLUZZI, SIMONE DIPACE, MARIA LUISA FIORELLA
Motor speech disorders and dysphagia in neurodegenerative diseases: clinical profiles compared 55

ALESSANDRA FERRERA, GLORIA GAGLIARDI
Acoustic biomarkers of Parkinson's Disease: a pilot study on the *Italian Parkinson's Voice and Speech* dataset 63

BARBARA GILI FIVELA, RAFFAELLO CASERTA, SONIA D'APOLITO, ANNA CHIARA PAGLIARO, VINCENZO SALLUSTIO, MARCO SINISCALCHI
Phonetic analysis and deep neural network examination for dysarthria detection beyond sociolinguistic differences: a comparison 83

CLAUDIO FERRANTE, VINCENZO SCOTTI
Cross-Lingual Transferability of Voice Analysis Models: a Parkinson's Disease Case Study 99

MARTA MAFFIA, LOREDANA SCHETTINO, ROSA DE MICCO, ALESSANDRO TESSITORE
Disfluencies in Parkinson's Disease. A study on Italian early-stage patients 117

VALENTINA COLONNA, FEDERICO LO IACONO, ANTONIO ROMANO
Amelia Rosselli e l'angoscia del respiro: tra stile e riflessi di parlato patologico 131

CLAUDIO ZMARICH, SIMONA BERNARDINI, SABRINA BONICHINI, MIRKO BONOTTO, FEDERICA CHIARI, ERIKA GANASSIN, CATERINA PISCIOTTA, FABIOLA VULLO, ELISA ZACHEO
Linguistic proficiency, communicative attitude, and severity in a sample of children who stutter aged 4-6 years 153

PARTE II
Parlato nelle demenze e nell'autismo

ELENA MARTINELLI, VITA GARRAMMONE, FRANCESCA MORI, INCORONATA NOLÈ, FRANCA CAMERIERO, MATILDE MARTINO, GAETANO DI BELLO, GLORIA GAGLIARDI
I correlati linguistici del deterioramento cognitivo: raccolta e analisi di un corpus di eloquio patologico di individui anziani lucani affetti da demenza 181

FRANCESCO SIGONA, BARBARA GILI FIVELA, PIETRO VIGORELLI, ANDREA BOLIOLI
A computational linguistic analysis of speech in elderly individuals with cognitive decline: the "Anchise-2022" 203

SYRIA CIRA IMPARATO, MARIA IZZO, OLGA LIGUORI, ERNESTO ORSINO, DONATA SENSALE, TINA SANTARPIA, GIULIA PERFETTO, GIOVANNA GISON, GLORIA GAGLIARDI
Caratteristiche pragmatiche dell'eloquio nel Disturbo dello Spettro Autistico (ASD): uno studio pilota su un campione di bambini italofoni residenti in Campania 229

VIRGINIA CHITI, VALENTINA SACCONE, ALESSANDRO PANUNZI
L'alterazione degli aspetti pragmatici e prosodici nel Disturbo dello Spettro Autistico: analisi del parlato di bambini in età scolare 251

PARTE III
Variazione e apprendimento della seconda lingua

DUCCIO PICCARDI
Reference, sound symbolism, and variation. In search of an elusive link 273

ANNA ANASTASENI, ANTONIO ROMANO, SYLVAIN GERBER
Approssimanti e laterali palatali in italiano: l'effetto della pronuncia sulla conversione fono-grafemica in un gruppo di giovani parlanti torinesi 299

ROBERTA BIANCA LUZIETTI
La degeminazione delle rotiche in Toscana: indagine preliminare in un archivio orale storico 317

GIUDITTA AVANO, CINZIA AVESANI, MARIO VAYRA
Meccanismi di livellamento in contesto universitario: la Gorgia Toscana in soggetti fiorentini fuorisede 337

KIM A. GROOTHUIS, GIOVANNI LEO, GIUSEPPE MAGISTRO, CLAUDIA CROCCO
Prosodic features of pragmatic *quello* in Informal Neapolitan Italian: data from real-world speech 361

VINCENZO SCOTTI, ROBERTO TEDESCO
On the Role of Dialogue Context in Predicting Speaking Style 381

FEDERICA CAVICCHIO, MIRKO GRIMALDI
I training audio e audiovisivi migliorano in misura uguale la percezione segmentale della seconda lingua 399

NICHOLAS NESE
Arabo lingua straniera. Acquisizione di nuove categorie fonologiche 415

BIANCA MARIA DE PAOLIS, ANTONIO ROMANO, CECILIA ANDORNO
Prosodic and syntactic markers of narrow focus in Italian L1 and L2 (French L1): an experimental study 437

PARTE IV
Riconoscimento automatico del parlato e del parlante

ELEONORA RODEGHER, ALESSIO BRUTTI, DANIELE FALAVIGNA
Automatic speech recognition for low-resource languages: exploring transfer learning from related languages 465

FRANCESCO SIGONA, MIRKO GRIMALDI
Calibration and Fusion of Parametric and Non-parametric Statistical Methods for Forensic Semi-Automatic Speaker Recognition Systems: experimental results 473

FRANCESCO SIGONA, GIUSEPPE VITOLO, MIRKO GRIMALDI
Forensic Automatic Speaker Recognition based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network 487

Autori 501

DOI: 10.17469/O2111AISV000001

BARBARA GILI FIVELA, SONIA D'APOLITO, ANNA CHIARA PAGLIARO,
VINCENZO SALLUSTIO, MARIA LUISA FIORELLA, MARCO SINISCALCHI

Introduzione

La lingua orale è un oggetto di studio ideale per ricavare informazioni sull'individuo, sul suo sistema sociolinguistico di riferimento, e quindi sulla sua identità linguistica, ma anche sullo stato complessivo dei sistemi fisiologici che gli permettono la produzione e percezione di parlato. L'analisi dell'eloquio offre informazioni preziose e lo sviluppo di applicazioni volte ad automatizzare tale analisi può avere forti ricadute anche in ambito socio-sanitario. L'aumento del numero di soggetti affetti da malattie neurodegenerative, come la malattia di Parkinson e la malattia di Alzheimer, che influiscono in vari modi sulla capacità di comunicare dell'individuo, suggerisce quale possa essere la ricaduta delle indagini interdisciplinari che facciano della voce e del parlato il loro obiettivo d'indagine.

In questo volume è raccolta una selezione di contributi presentati durante il convegno dal titolo "Il parlato in ambito medico: analisi linguistica, applicazioni tecnologiche e strumenti clinici/ Spoken language in the medical field: linguistic analysis, technological applications and clinical tools", svoltosi a Lecce dal 15 al 17 febbraio 2023. Il convegno è stato organizzato prendendo spunto dal progetto PRIN 2017JNKCYZ "Phonetic analysis of dysarthric speech by speakers of different varieties of Italian to develop clinical tools" e ha rappresentato un'occasione di discussione tra studiosi di settori scientifici diversi che affrontano in modo multi- e interdisciplinare lo studio del parlato, con attenzione alle applicazioni tecnologiche in ambito clinico.

I contributi presenti nel volume riguardano vari temi relativi sia al parlato normofasico che al parlato patologico, e mostrano che l'analisi dell'eloquio, automatica o meno, offre informazioni importanti sia sul piano teorico che su quello applicativo.

Il volume, pubblicato grazie ai fondi del PRIN summenzionato, si apre con il contributo di Isabella Laura Simone "Language Impairment in Neurodegenerative diseases", che descrive in modo molto chiaro le basi neurofisiologiche del linguaggio e quali sono gli aspetti del linguaggio e del parlato che possono essere considerati marcatori di sindromi specifiche o di malattie neurodegenerative. Nell'articolo di Serge Pinto "Speech motor control from its impairment: the relevance of studying dysarthria in Parkinson's disease" l'attenzione è rivolta alle disartrie e, in particolare, alla disartria ipocinetica tipica della malattia di Parkinson. Si discute dell'importanza dello studio della dimensione universale e di quella linguo-specifica, perché l'attenzione a questi aspetti può aiutare sia a sviluppare terapie personalizzate, sia a

comprendere ulteriormente il ruolo di alcune strutture neurali nel controllo motorio esercitato durante il parlato.

Nella sezione "Disturbi del parlato e della fluenza", sono raccolti contributi multi- e interdisciplinari, basati sia su analisi manuali che su elaborazioni computazionali. Diletta Sciancalepore, Valentina Lavermicocca, Tresanna Illuzzi, Laura Lumaca, Simone Dipace e Maria Luisa Fiorella firmano l'articolo dal titolo "Motor speech disorders and dysphagia in neurodegenerative disesases: clinical profiles compared". Gli autori considerano diverse malattie neurodegenerative e analizzano i principali correlati acustici delle alterazioni nella fonazione e articolazione del parlato nonché le difficoltà nella deglutizione, concludendo che, nel caso disartria e disfagia coesistano, la prima è normalmente peggiore della seconda. Nel contributo "Acoustic biomarkers of Parkinson's Disease: a pilot study on the *Italian Parkinson's Voice and Speech* dataset", Alessandra Ferrera e Gloria Gagliardi presentano i risultati preliminari di un'analisi acustica del parlato di 65 soggetti dell'Italia meridionale, tra i quali 28 pazienti Parkinson, confermando l'alterazione di vari parametri acustici nei Parkinson rispetto ai soggetti di controllo e riportando la tendenza al cambiamento delle misure di durata a seconda del genere del parlante.

Nell'articolo dal titolo "Phonetic analysis and deep neural network examination for dysarthria detection beyond sociolinguistic differences: a comparison", Barbara Gili Fivela, Raffaello Caserta, Sonia d'Apolito, Anna Chiara Pagliaro, Vincenzo Sallustio e Marco Siniscalchi confrontano l'analisi fonetica e l'uso di *deep learning model* nella classificazione di parlato disartrico e normotipico. L'analisi degli stessi materiali con metodi di fonetica acustica e tramite una rete neurale (una *Convolutional Neural Network* addestrata con dati in lingua spagnola e, tramite *transfer learning*, oggetto di *fine tuning* grazie a dati in italiano) suggerisce che i diversi metodi possano essere integrati con successo. Claudio Ferrante e Vincenzo Scotti, in "Cross-lingual Transferability of Voice Analysis Models: a Parkinson's Disease Case Study", propongono nuovamente l'uso dei *deep learning model* per estrarre *feature* acustiche e ottengono buoni risultati nel caso dell'individuazione del Parkinson grazie al transfer cross-linguistico e all'adattamento di modelli dell'inglese al telugu.

Sempre in merito al Parkinson, nell'articolo "Disfluencies in Parkinson's Disease. A study on Italian early-stage patients", Marta Maffia, Loredana Schettino, Rosa de Micco e Alessandro Tessitore discutono dati relativi alle disfluenze in parlato monologico, suggerendo che esistano tipi, caratteristiche e usi specifici delle disfluenze già nei primi stadi della malattia. L'articolo successivo si intitola "Amelia Rosselli e l'angoscia del respiro: tra stile e riflessi di parlato patologico" ed è firmato da Valentina Colonna, Federico Lo Iacono e Antonio Romano. Il contributo riguarda l'analisi delle letture della poetessa Amalia Rosselli tra il 1985 e il 1990. Lo stile prosodico della poetessa viene messo in relazione con le informazioni relative al suo stato di salute in quel periodo.

Sulla balbuzie in bambini di età prescolare si concentrano Claudio Zmarich, Simona Bernardini, Sabrina Bonichini, Mirko Bonotto, Federica Chiari, Erika

Ganassin, Caterina Pisciotta, Fabiola Vullo ed Elisa Zacheo nell'articolo "Linguistic proficiency, communicative attitude and severity in a sample of pre-school children who stutter". Discutendo una serie di esperimenti, gli autori riportano che non sussistono differenze significative nella *communication attitude* dei soggetti sperimentali, mentre risultati peggiori si riscontrano nei test di articolazione e ripetizione dei balbuzienti rispetto a quelli del gruppo di controllo.

La sezione "Parlato nelle demenze e nell'autismo" si apre con l'articolo di Elena Martinelli, Vita Garramone, Francesca Mori, Incoronata Nolè, Franca Cameriero, Matilde Martino, Gaetano Di Bello e Gloria Gagliardi, dal titolo "I correlati linguistici del deterioramento cognitivo: raccolta e analisi di un corpus di eloquio patologico di individui anziani lucani affetti da demenza". Lo studio riguarda 20 soggetti affetti da vari tipi di demenza e 20 soggetti di controllo. L'analisi di 9 ore di parlato semispontaneo, effettuata tramite test statistici inferenziali, evidenzia il deterioramento linguistico a livello fonetico, morfosintattico/lessicale, sintattico e semantico, mentre l'analisi qualitativa permette di individuare la carenza di coerenza e coesione, nonché la presenza di ripetizioni e strategie di compensazione. Riguarda le demenze anche l'articolo di Francesco Sigona, Barbara Gili Fivela, Pietro Vigorelli e Andrea Bolioli, dal titolo "A computational linguistic analysis of speech in elderly individuals with cognitive decline: the Anchise-2022". L'analisi computazionale di circa 200 conversazioni tra *caregiver* e anziani con declino cognitivo è condotta con tecniche di Natural Language Processing (NLP) statistiche descrittive e analisi inferenziali, che vengono correlate con dati sulla progressione della demenza misurata con Mini-Mental State Examination (MMSE).

Nell'articolo "Caratteristiche pragmatiche dell'eloquio nel Disturbo dello Spettro Autistico (ASD): uno studio pilota su un campione di bambini italofoni residenti in Campania", Syria Cira Imparato, Maria Izzo, Olga Liguori, Ernesto Orsino, Donata Sensale, Tina Santarpia, Giulia Perfetto, Giovanna Gison e Gloria Gagliardi esaminano le caratteristiche pragmatiche e conversazionali del parlato semi-spontaneo prodotto da 17 bambini in età scolare con diagnosi di Autismo High-Functioning e 17 con sviluppo tipico. I testi prodotti dai bambini ASD risultano meno articolati e più frammentanti internamente. Un minor numero di connettivi e pause piene confermano il deficit pragmatico, ma suggeriscono anche interessanti interazioni tra pragmatica, morfosintassi e fluenza nel parlato. La pragmatica nel parlato di soggetti con ASD è discussa anche nel contributo di Virginia Chiti, Valentina Saccone e Alessandro Panunzi, dal titolo "L'alterazione degli aspetti pragmatici e prosodici nel Disturbo dello Spettro Autistico: analisi del parlato di bambini in età scolare". Gli autori descrivono l'analisi delle caratteristiche acustiche, lessicali e di struttura informativa del parlato semispontaneo di 9 bambini con ASD (8 e i 12 anni), mettendo in evidenza la variazione presente tra parlato monotono ed estremamente enfatico, nonché la frammentazione.

La sezione su "Variazione e apprendimento della seconda lingua" si apre con il contributo di Duccio Piccardi "Reference, sound symbolism, and variation. In search of an elusive link", che riguarda il fonosimbolismo e la proposta dell'esistenza

del *frequency code,* da parte di Ohala, attraverso lo studio della possibile relazione tra la durata dell'aspirazione delle occlusive sorde in fiorentino e la percezione di virilità. Lo studio delle risposte in un test a risposta forzata non conferma la relazione ipotizzata, anche se i risultati sono compatibili con l'esistenza di un transfer di aspetti fonosimbolici tra diversi domini di significato. Nell'articolo "Approssimanti e laterali palatali in italiano: l'effetto della pronuncia sulla conversione fono-grafemica in un gruppo di giovani parlanti torinesi", Anna Anastaseni, Antonio Romano e Sylvain Gerber si occupano della neutralizzazione dell'opposizione /ʎ/-/j/ a favore di foni [j]-like da parte di alcuni parlanti. L'analisi di dati di percezione e acustici mette in evidenza il ruolo della durata nell'identificazione dei fonemi indagati e permette di individuare un gruppo di parlanti che neutralizza la distinzione e un altro che, invece, mantiene la differenza. Roberta Bianca Luzietti propone, invece, un articolo su "La degeminazione delle rotiche in Toscana: indagine preliminare in un archivio orale storico", nel quale descrive i risultati preliminari di un'indagine sociofonetica sulla degeminazione nella varietà di Prato, "fotografata" nell'archivio orale di Angela Spinelli. L'archivio è particolarmente prezioso, perché permette l'analisi di parlato non-standard e più spontaneo di quello di laboratorio, tipicamente usato per indagini acustiche.

Giuditta Avano, Cinzia Avesani, Mario Vayra firmano un articolo dal titolo "Meccanismi di livellamento in contesto universitario: la Gorgia Toscana in soggetti fiorentini fuorisede", nel quale studiano la spirantizzazione delle occlusive in posizione intervocalica da parte di fiorentini che hanno studiato a Bologna. Il confronto con studenti fiorentini che non si sono spostati a Bologna offre l'occasione per considerare vari fattori linguistici e sociolinguistici e osservare un minor numero di casi di spirantizzazione negli studenti fuorisede, spia di un fenomeno di *levelling* che interessa la Gorgia Toscana. Tuttavia, il fenomeno sembra quasi inconscio e fortemente influenzato da aspetti sociolinguistici, differendo da quello precedentemente osservato in lavoratori fiorentini che si sono spostati a Bologna. Kim Groothuis, Giovanni Leo, Giuseppe Magistro e Claudia Crocco presentano un contributo sulle "Prosodic features of pragmatic *quello* in Informal Neapolitan Italian: data from real-world speech", nel quale utilizzano un corpus di parlato raccolto nel canale YouTube con il software Pyrlato. Nell'articolo, gli autori indagano le caratteristiche prosodiche del pronome dimostrativo *quello* che, in Campania, presenta anche usi pragmatici. L'analisi delle 248 istanze rilevanti per l'indagine (tra le 1775 occorrenze estratte da YouTube grazie a Pyrlato) mostra che l'uso pragmatico di *quello* corrisponde alla presenza di un accento intonativo e di un confine prosodico. Gli autori suggeriscono anche un'ipotesi sull'origine dell'uso pragmatico di *quello* e del corrispettivo dialettale.

Sul versante delle tecnologie del parlato, Vincenzo Scotti e Roberto Tedesco discutono di sintesi da testo scritto nell'articolo "On the Role of Dialogue Context in Predicting Speaking Style". Gli autori prendono in considerazione l'efficacia di modelli recenti, che permettono di predire lo stile d'eloquio dal testo della conversazione, pensando alle ricadute per agenti conversazionali e chatbot. In particolare,

descrivono lo sviluppo e addestramento di una rete neurale che, per predire lo stile d'eloquio, funge da collegamento tra la componente testuale e quella di sintesi vocale di un chatbot.

Per quanto riguarda l'apprendimento della seconda lingua, Federica Cavicchio e Mirko Grimaldi affrontano il tema "I training audio e audiovisivi migliorano in misura uguale la percezione segmentale della seconda lingua", verificando l'efficacia dell'*High Variability Phonetic Training*, effettuato usando materiali audiovideo, per migliorare la percezione dell'inglese L2 nel caso di italofoni. L'analisi del post-test dei gruppi sperimentali, che hanno effettuato un addestramento solo audio oppure audio-video, e del gruppo di controllo suggerisce che l'accuratezza nell'identificazione di fonemi migliora con qualsiasi addestramento e che le informazioni audiovisive influenzano l'identificazione delle vocali L2, seppur temporaneamente. Gli autori concludono che, in linea con il modello *Analysis by Synthesis,* le rappresentazioni fonetico-fonologiche si basano principalmente sulle caratteristiche acustiche del segnale. Nicolas Nese, invece, affronta la questione "Arabo lingua straniera. Acquisizione di nuove categorie fonologiche", orientando l'attenzione sull'acquisizione di foni consonantici dell'arabo da parte di 64 studenti italofoni, in particolare l'acquisizione delle faringalizzate [tʕ], [dʕ], [sʕ], [ðʕ], delle faringali [ħ], [ʕ], e dell'uvulare [q]. I risultati di un compito di lettura, uno di identificazione e uno di discriminazione migliorano in base alla competenza dell'apprendente, mostrando una differenziazione dei foni in termini di difficoltà.

Sempre in questa sezione, Bianca Maria De Paolis, Antonio Romano e Cecilia Andorno presentano un contributo dal titolo "Prosodic and syntactic markers of narrow focus in Italian L1 and L2 (French L1): an experimental study", nel quale analizzano le strategie per la realizzazione del *narrow focus*, correttivo e non, in 41 parlanti italiani e francesi, questi ultimi registrati mentre parlano la L1 e l'italiano L2. Influenze crosslinguistiche emergono sia a livello sintattico che prosodico. Le somiglianze tra i due codici inizialmente aiutano nella segnalazione del *narrow focus* tramite l'uso delle frasi scisse, mentre le differenze prosodiche non favoriscono la segnalazione grazie all'uso della prosodia.

Infine, tre lavori compongono la sezione inerente ad applicazioni tecnologie per il "Riconoscimento automatico del parlato e del parlante", alcune delle quali riguardano l'ambito forense, e non quello medico, ma che possono fornire ulteriori spunti per l'analisi del parlato. Eleonora Rodegher, Alessio Brutti e Daniele Falavigna, con "Automatic speech recognition for low-resource languages: exploring transfer learning from related languages", descrivono alcuni esperimenti relativi all'uso della tecnica del *transfer learning*. Grazie a studi che riguardano italiano, spagnolo, portoghese e tedesco, così come galiziano e vallader, gli autori verificano che il *fine-tuning,* effettuato per adattare ad una specifica lingua o varietà di lingua un modello sviluppato grazie a molti dati relativi ad un'altra lingua, è più funzionale in termini di *error rate* rispetto all'uso di un modello addestrato su pochi dati.

Francesco Sigona e Mirko Grimaldi, con "Calibration and Fusion of Parametric and Non-parametric Statistical Methods for Forensic Semi-Automatic Speaker

Recognition Systems: experimental results" prendono in esame l'accuratezza degli approcci più usati in Italia per il riconoscimento automatico del parlante in ambito forense, emulati nel software IMPAVIDO. L'ipotesi è che la calibrazione della regressione lineare migliori l'accuratezza. La metrica usata per stabilire l'accuratezza è il *likelihood-ratio cost* e i fattori considerati sono molteplici, inclusa la dimensione del database e il modo in cui vengono considerate le caratteristiche disponibili per le vocali. Infine, Francesco Sigona, Giuseppe Vitolo e Mirko Grimaldi si occupano di comparazione automatica della voce in ambito forense, in "Forensic Automatic Speaker Recognition based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network". Dopo aver discusso le differenze tra applicazioni in ambito forense e non, gli autori riportano i confortanti risultati preliminari ottenuti con l'uso dell'*Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network*, che migliorano all'aumentare della durata dei campioni di parlato e alla diminuzione del rumore di fondo.

I curatori ringraziano sentitamente i membri del Comitato Scientifico del convegno svoltosi a Lecce e tutti gli studiosi che hanno agito come revisori degli *abstract* inviati in occasione del convegno e dei contributi pubblicati in questo volume: Giovanni Abete (Università degli Studi di Napoli "Federico II"), Paola Angelelli (Università del Salento), Cinzia Avesani (ISTC-CNR Padova), Valentina Bambini (Scuola Universitaria Superiore IUSS Pavia), Sarah Bigi (Università Cattolica del Sacro Cuore), Silvia Calamai (Università di Siena), Chiara Celata (Università di Urbino "Carlo Bo"), Sonia Cenceschi (Scuola Universitaria Professionale della Svizzera Italiana), Claudia Crocco (Università di Gent, Belgio), Francesco Cutugno (Università degli Studi di Napoli "Federico II"), Valentina De Iacovo (Università di Torino), Anna De Meo (Università degli Studi di Napoli L'Orientale), Francesca Dovetto (Università degli Studi di Napoli "Federico II"), Mariaelena Favilla (Università di Modena e Reggio Emilia), Vincenzo Galatà (ISTC-CNR Padova), Gloria Gagliardi (Università di Bologna), Giovanni Gobber (Università Cattolica del Sacro Cuore), Paola Leone (Università del Salento), Robin Lickley (Queen Margaret University), Paolo Mairano (Université de Lille), Giovanna Marotta (Università di Pisa), Pietro Maturi (Università degli Studi di Napoli "Federico II"), Chiara Meluzzi (Università di Milano), Rosalba Nodari (Università degli Studi di Siena), Antonio Origlia (Università degli Studi di Napoli "Federico II"), Maria Palmerini (Cedat85), Alessandro Panunzi (Università di Firenze), Elisa Pellegrino (Universität Zürich, Svizzera), Tommaso Raso (Universidade Federal de Minas Gerais, Brasile), Antonio Romano (Università di Torino), Michela Russo (University Jean-Moulin Lyon 3, Francia), Stephan Schmid (Universität Zürich, Svizzera), Francesco Sigona (Università del Salento), Patrizia Sorianello (Università di Bari), Lorenzo Spreafico (Libera Università di Bolzano), Mario Vayra (Università di Bologna), Alessandro Vietti (Libera Università di Bolzano), Claudio Zmarich (ISTC-CNR Padova).

DOI: 10.17469/O2111AISV000002

ISABELLA LAURA SIMONE

Language impairment in neurodegenerative diseases

Language and speech are among of the most complex activities, involving cognitive-linguistic processes, motor speech planning, programming control and neuromuscular execution. Two specific brain areas generate speech: -frontal Broca's area related to speech production and articulation, and -temporal Wernicke's area responsible for speech comprehension. However, organization of language and comprehension/production of speech are mediated by broader neural network covering several cortical-subcortical regions and their connections. Some disorders of nervous system show up with language/speech changes. Language in neurodegenerative diseases may represent a specific marker to distinguish language isolated syndromes, as primary progressive aphasia (PPA) and its variants, but also to diagnose neurodegenerative disorders characterized by language impairment within other cognitive and neurological disorders, as Parkinson, Alzheimer, Amyotrophic Lateral Sclerosis, Multiple Sclerosis. The study of language network may contribute to greater understanding of extension of neurodegenerative processes. Our article is a review on language impairments that characterize the most frequent neurodegenerative diseases.

Keywords: language, speech, neurodegenerative diseases, cognitive functions.

1. *Introduction*

"Language" and "speech" are among the most complex human activities involving cognitive-linguistic processes, speech motor planning, programming check, and neuromuscular execution.

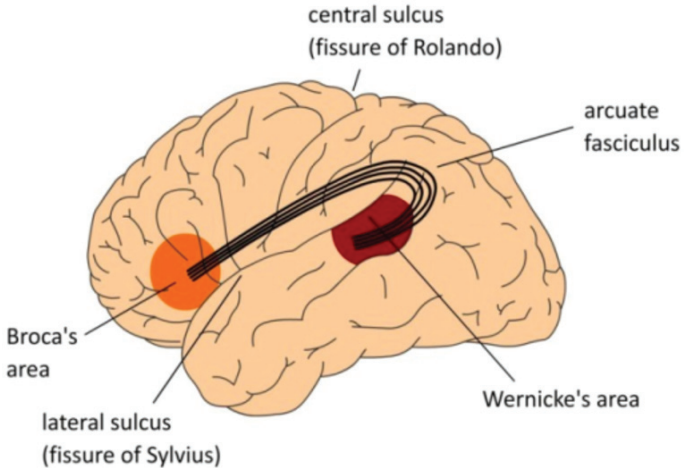
"**Language**" is the cognitive ability to communicate, to assign or decode symbols and to organize or decipher sentences and phrases. In this definition, language is the "*engine*" of communication. When language is impaired due to injury to the brain, the disorder is referred to "*aphasia*".

"**Speech**" is a motor ability of speaking, involving the lungs, trachea, respiratory muscles, vocal cords, mouth, tongue and velum, facial muscles and is defined as the mechanism by which language is expressed and articulated orally. Some individuals may have a motor speech deficit, making it difficult to produce words, e.g. "*dysarthria*", or to coordinate complex movements for articulation, e.g. "*speech apraxia*".

In 1874, Carl Wernicke proposed the first brain behavioral model, named Wernicke-Geschwind model, based on the idea that language consists of two basic functions: *a)* language comprehension which is a sensory/perceptual function, and *b)* language production, which is a motor function. In this simple model, the language function was localized in two main regions: (1) an area sited in the *left* temporal lobe (known as *Wernicke's area*) that contains auditory images of words

responsible for speech perception. A lesion at this level causes "*sensory or fluent aphasia*"; (2) an area located in the *left* inferior frontal lobe (known as *Broca's area*) that contains the motor memories responsible for speech production. A lesion at this level causes "*motor aphasia or nonfluent aphasia*". These two regions communicate with each other through a bundle of fibers, identified as the *arcuate fasciclus*, essential for language, from spontaneous speech and word retrieval to repetition and comprehension skills. Lesion of the arcuate fasciculus can result in *conduction aphasia* as the inability to repeat words or phrases (Figure 1).

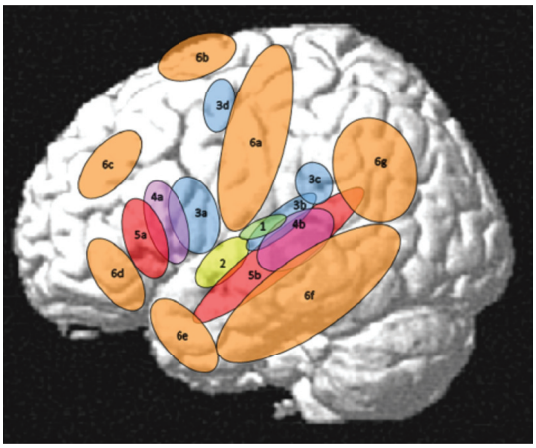
Figure 1 - *Wernicke-Geschwind model (from Hagoort P. 2013, n. 22)*



However, this simple model is obsolete, and it is known that language involves several distinct but integrated areas located in a vast network of many brain regions. These brain regions are represented by a "core" and a number of "margin regions" (Hertrich, Dietrich & Ackermann, 2020).

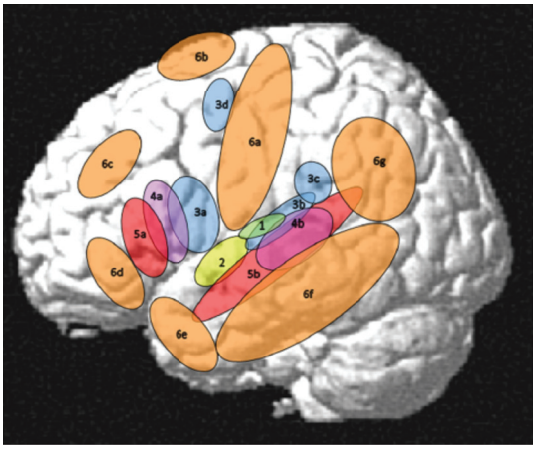
The "**core**" language network is restricted to extensive areas of the ***left*** temporal and frontal lobes including: 1) the auditory cortex as the primary input structure; 2) the auditory area of word form as the central region of comprehension of the words, – "Wernicke's area"; 3) the posterior part of the frontal language area, – "Broca area" and the posterior superior temporal gyrus as phonological regions that link an auditory phonetic to an articulatory linguistic code; 4) the middle part of the frontal language areas and the superior and middle temporal gyrus as for syntax processing, manipulation and detection of structures above the word level; 5) the anterior part of the frontal language areas, anterior and posterior parts of superior temporal gyrus as fundamental lexical-semantic areas linking phonological codes to lexical meanings (Hertrich *et al.* 2020) (Figure 2a).

Figure 2a - *Core language network. (1) Auditory cortex; (2) Auditory area of word form as the central region of comprehension of the words; (3 a,b,c,d) Phonological areas linking an auditory-phonetic to an articulatory language code; (4 a,b,) Syntax processing, manipulating and detecting structures above word level; (5a,b) Lexical-semantic core areas linking phonological codes to lexical meanings (from Hertrich et al. 2020, Front Commun, n.23)*



The "**margin regions**" are other **bilateral** brain structures and networks that complete the basic linguistic system, represented by 1) parietal and posterior temporal regions contributing to sensorimotor processing, 2) inferior temporal-occipital regions predominantly linked to visual object representations and their association with language processing, and 3) the temporal pole as a language interface toward the processing of emotions, valence, and social cognition (Hertrich *et al.* 2020) (Figure 2b).

Figure 2b - *Margin language network. Margin regions: (6a) Sensorimotor cortex; (6b) Supplementary motor area (SMA) and pre-SMA; (6c) dorsolateral prefrontal cortex; (6d) orbitofrontal cortex; (6e) temporal pole; (6f) middle and inferior temporal regions; (6g) parietal and temporoparietal regions (from Hertrich et al. 2020, Front Commun, n.23)*



Finally, subcortical structures are also involved in the production of language, as the thalamus, caudate nucleus, globus pallidus, subthalamic nucleus, substantia nigra and cerebellum, which on the one hand regulate and coordinate the motor aspects of language, on the other hand play a role in the processing of language-cognitive functions.

Specifically, the cerebellum exerts control over speech and language functions in several ways. The most obvious is the regulatory motor control, which is a major function of the cerebellum, *-ataxic dysarthria* resulting from injury. In addition, several neuroanatomic, clinical, neurophysiological and neuroimaging evidences demonstrated the crucial role of the cerebellum in cognitive processing and language function. Functional MRI studies have shown that cerebellum has reciprocal connections with both left inferior frontal gyrus and left lateral temporal cortex (Booth, Wood, Lu, Houk & Bitan, 2007), as well as associations between neural activity in the posterolateral cerebellar and cortical activity of regions related to social cognition, memory, language and executive functions (Highnam, Bleile, 2011). Scientific literature defines this concept as "*cerebellar neurocognition*" (Baillieux, De Smet, Paquier, DeDeyn & Marien, 2008).

Similarly, the basal ganglia and the connected thalamus may control speech and language in several ways. The basal ganglia have a function in motor processing, including articulation, but they are also involved in the cognitive control required by linguistic and non-linguistic processing (Nadeau, 2021; Booth *et al.*, 2007).

The thalamus moderates the transfer of lexical information to cortical areas. Indeed, a thalamic lesion can lead to a loss of spontaneous speech due to the interruption of the excitatory input from the thalamus to the cortex. (Nadeau, Crosson B, 1997)

Neurodegenerative diseases are a large group of disorders affecting the nervous system and represent a major common cause of disability worldwide. One of the most difficult challenges for individuals with neurodegenerative disease is the insidious onset of communication impairments.

Our article is a review on language impairments that characterize the most frequent neurodegenerative diseases.

2. *Neurodegenerative diseases and language impairments*

Neurodegenerative diseases are a heterogeneous group of disorders characterized by the progressive degeneration of the structures and functions of the central nervous system. Several factors contribute to neurodegenerative diseases, including genetic mutations, neuroinflammation, mitochondrial dysfunction, neuronal cell death, misfolded protein aggregation, i.e. amyloid-β (Aβ) and tau in Alzheimer's Disease, mutant huntingtin (mHTT) in Huntington's chorea, α-synuclein in Parkinson's Disease, and TAR DNA-binding protein-43 (TDP-43), fused in sarcoma/translocated in liposarcoma (FUS/TLS) in Amyotrophic lateral sclerosis, (Hou, Dan, Babbar, Wei, Hasselbalch, Croteau & Bohr, 2019; Peng, Trojanowski & Lee, 2020).

Neurodegenerative diseases share some pathological and clinical signs related to the atrophy of the brain area connected to specific functions, such as cognitive functions with memory loss, disorientation, language impairment.

The impairment of language or speech functions occuring in neurodegenerative diseases could be attributed to disruptions in linguistic processes (i.e., related to conceptualization, word-finding, syntax, semantics, and/or phonology), or to motor processes involved in the planning and programming (i.e., apraxia of speech) or executing speech motor commands (i.e., dysarthria), or any combination of these processes. Language deficits are present in several neurodegenerative diseases, sometimes in the early stages, as a selective and prominent symptom, as in Primary Progressive Aphasia (PPA), or in combination with other cognitive and neurological disorders, as in Alzheimer's disease (AD), or movements disorders.

Our article is a review on the most important and common neurodegenerative diseases as Parkinson's disease (PD), Alzheimer's disease (AD), Frontotemporal dementia (FTD), Amyotrophic Lateral Sclerosis (ALS), and multiple sclerosis, all showing complex and hetereogenous impairment of language functions.

2.1 Parkinson's disease

Parkinson's disease (PD) is a chronic neurodegenerative disease characterized by the progressive loss of dopaminergic neurons, mainly in the pars compacta of substantia nigra (Damier, Hirsch, Agid & Graybiel, 1999; Redgrave, Rodriguez, Smith, Rodriguez-Oroz, Lehericy, Bergman & Obeso, 2010). Furthermore, there is some evidence that caudal, brainstem and other non-dopaminergic neurons are also affected. Speech and language impairments have been described in approximately 90 % of patients with PD. In particular, the impairment of motor function in PD may involve also the motor speech control, with difficulty in breathing, phonation, articulation, resonance and prosody (Goberman, Coelho, 2002). Patients with PD may have low volume (hypophonia), impaired voice quality (dysphonia), reduced range of articulatory movements (hypokinetic) and reduced pitch/volume inflection to convey meaning or emotion (hypoprosodia). *"Hypokinetic dysarthria is the generic term for this set of defects"*. Hypokinetic dysarthria worsens with disease progression, is closely related to overall intelligibility, and becomes very unresponsive to pharmacological treatments. The haste (words rushing together quickly) and disfluency (from hesitation to loss of pace), the tendency to accelerate emission of sounds, the overlapping syllables and omitting the end of words may further impair understanding. Echolalia and palilalia may characterize the language features in more advanced stages of the disease, and the degree of communication impairment can progress to inaudible and unintelligible speech. In addition to the above-described language impairment, patients with PD may have subtle deficits in processing and comprehension of verbal and non-verbal emotional aspects of communication related to cognitive-linguistic dysfunction (Grossman, Crino, Reivich, Stern & Hurtig, 1992; Grossman, Kalmanson, Bernhardt, Morris, Stern & Hurtig, 2000).

2.2 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease that affects millions of people worldwide and is mostly recognizable through its primarily cognitive syndrome: the dementia. It is a progressive neurological disease that results in irreversible loss of neurons, in particular in the cortex and hippocampus, which leads to characteristic memory loss and behavioral changes (McKhann, Drachman, Folstein, Katzman, Price & Stadlan, 1984).

AD is characterized by aberrant accumulation of beta-amyloid plaques and tau protein tangles with consequent disruption in brain networks. A recent study reported a relationship between beta-amyloid deposition and brain regions involved in memory, visual-spatial, and language functions (Baboo, Prasad, Mahajan & Baths, 2022).

AD dementia is associated with heterogeneous and broad cognitive impairment; the most typical hallmarks are impairment of memory, executive function, and language (Vigo, Coelho & Reis, 2022). In AD, language and memory functions are closely related and difficulties in productive language, speech understanding, and memory impairment overlap.

Language impairments have been reported to occur in preclinical AD, as well as in mild, moderate, and severe AD dementia (Bayles, Tomoeda, Cruz & Mahendra, 2000). In the early stage of AD, many communication skills remain intact, including speech sound production, grammar, and conversational abilities; the patients may show deficits in word finding, comprehending complex conversation and memory lapses. In the middle stage, several deficits are found in word finding, understanding of complex tasks, reading comprehension, and impairment in many domains of memory. The linguistic deficit mainly occurs with a decline in lexical semantic abilities, word comprehension and impaired verbal fluency. (Forbes-McKay, Venneri, 2005; Taler, Phillips, 2008; Catricalà, Della Rosa, Plebani, Perani, Garrard & Cappa, 2015). In the late stage, patients show very limited expressive language, inappropriate word choice, echolalia and verbal stereotypies (Ferris, Farlow, 2013), up to mutism, severe deficits in auditory comprehension and memory deficits in all domains.

2.3 Frontotemporal dementia (FTD)

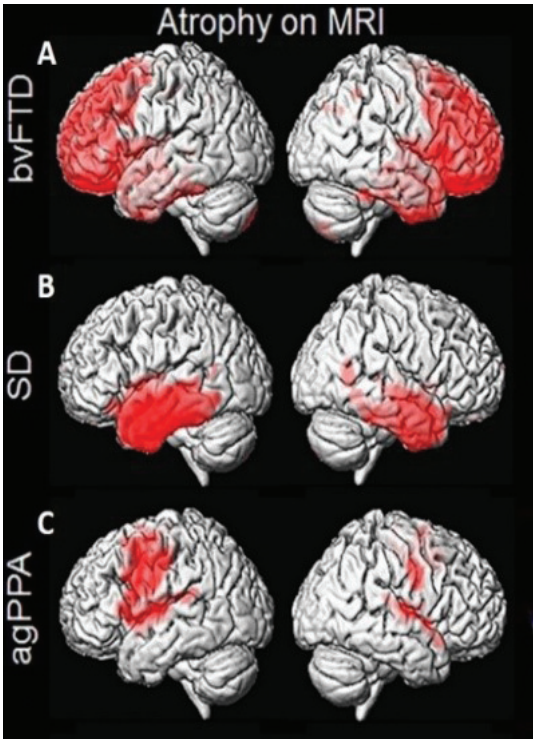
Frontotemporal dementia (FTD) is a group of clinically, pathologically and genetically heterogeneous disorders that primarily affect the left frontal and temporal lobes of the brain. The FTD is clinically characterized by progressive impairment of executive functions, changes in behavior, and a decrease in language proficiency (Bang, Spina & Miller, 2015). The most common genetic causes are mutations in progranulin, or the microtubule-associated protein tau, or a hexanucleotide repeat expansion in C9ORF72. Clinical phenotypes vary across the different genetic mutations.

The FTD is classified into four distinct core clinical syndromes, including behavioral-variant FTD (bvFTD) and three varieties of primary progressive aphasia (PPA), a) the nonfluent/agrammatic-variant primary progressive aphasia (nfvPPA),

b) the semantic-variant primary progressive aphasia (svPPA), and c) the logopenic variant PPA (lv-PPA):

– The *Behavioral variant frontotemporal dementia* (bvFTD) is the most common frontotemporal disorder that involves changes in personality, behavior and judgment, up to in awkward social situations. Patients are often unaware of their inappropriate behaviors (Rascovsky, Hodges, Knopman, Mendez, Kramer, Neuhaus, van Swieten, Seelaar, Dopper, Onyike, Hillis, Josephs, Boeve, Kertesz, Seeley, Rankin, Johnson, Gorno-Tempini, Rosen, ... & Miller, 2011). Over time, also speech and/or motor problems may occur. Studies of structural imaging show atrophy in frontal regions (orbitofrontal, middle frontal gyrus, rostromedial prefrontal cortex, pre-supplementary motor cortex), insula (dorsal and ventral anterior insula extending to posterior insula at late stages), anterior/mid cingulate cortex, anterior temporal lobes and subcortical structures (basal ganglia, thalamus, hippocampus), as well as cerebellum (Whitwell, 2019) (Figure 3-**A**).

Figure 3 - *Pattern of atrophy in FTD clinical phenotypes.* ***A*** *in bVFTD: atrophy in frontal, insula, anterior/middle cingulate cortex, anterior temporal lobes and subcortical structures: basal ganglia, thalamus, hippocampus, and cerebellum;* ***B*** *in svPPA: atrophy in anterior temporal pole (left>right hemisphere), frontoinsula, left middle and inferior temporal gyri, fusiform gyri, amygdala;* ***C*** *in nfvPPA: atrophy in inferior frontal gyrus, dorsolateral prefrontal cortex, supramarginal gyrus and the parietal lobes (Image adapted from Whitwell J.L. 2019, Prog Mol Biol Transl Sci., n.51)*



– The *Primary progressive aphasia (PPA*) involves changes in the ability to communicate, use language to speak, read, write, and understand what others are saying. Problems with memory, reasoning, and judgment as well as behavioral changes may develop with disease progression (Gorno-Tempini, Hillis, Weintraub, Kertesz, Mendez, Cappa, Ogar, Rohrer, Black, Boeve, Manes, Dronkers, Vandenberghe, Rascovsky, Patterson, Miller, Knopman, Hodges, Mesulam & Grossman, 2011) As above reported, the PPA includes three varieties according to the type of language problems encountered:
  a. Semantic variant PPA (svPPA) (previously known as semantic dementia) is characterized by loss of ability to understand single words, or the meaning of words, or difficulty naming objects. Brain atrophy in svPPA has been reported in the anterior temporal pole (left >right hemisphere), frontoinsula, left middle and inferior temporal gyri, fusiform gyrus, amygdala and basal forebrain (Figure 3-**B**) (Whitwell, 2019);
  b. Agrammatic PPA (previously known as progressive non-fluent aphasia) (nfvPPA) is characterized by errors and distortions in the sound of speech and/or agrammatism in speech (speech apraxia). In nfvPPA, greater progressive atrophy is found in inferior frontal gyrus, dorsolateral prefrontal cortex, supramarginal gyrus of the left hemisphere and the parietal lobes (Figure 3-**C**) (Whitwell, 2019);
  c. Logopenic PPA is characterized by difficulty to find the right words, pauses in speech, difficulty to repeat phrases or sentences. In logopenic PPA, atrophy of left temporoparietal and posterior cingulate is found.

2.4 Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by diffuse upper and lower motor neuron degeneration in four districts of nervous system: bulbar, cervical, thoracic, and lumbosacral, and consequent impairment of all motor functions also including respiratory, phonatory, swallowing functions (Mitchell, Borasio, 2007).

Approximately 80 to 96 % of people with ALS will become unable to speech during the disease progression (Kuhnlein, Gdynia, Sperfeld, Lindner-Pfleghar, Ludolph, Prosiegel & Riecke, 2008).

Due to motor dysfunction, ALS patients may have a progressive mixed flaccid-spastic dysarthria that is characterized by impaired articulation, slowed speech, difficulty in producing sentences, inappropriate pauses, hypernasality, hypophonia, rough or breathy voice quality, fatigue or shortness of breath with speech, reduced utterance length due to impaired breath support, strangled voice (Kuhnlein *et al.*, 2008). The voice impairment is one of the most important aspects of the bulbar dysfunction.

However, ALS is a multi-system disorder, impacting not only motor functions but also cognitive aspects, ranging from mild impairment, in ~50 % of ALS patients, to frontotemporal dementia in ~15 % of patients (Goldstein, Abrahams,

2013; Lomen-Hoerth, Murphy, Langmore, Kramer, Olney & Miller,2003; Neary, Snowden & Mann, 2000; Celsis, Argall, Henderson, McCombe & Robinson, 2020), with language involved up to semantic dementia or PPA (Taylor, Brown, Tsermentseli, Al-Chalabi, Shaw, Ellis, Leigh & Goldstein, 2013; Ringholz, Appel, Bradshaw, Cooke, Mosnik & Schulz, 2005; Strong, Abrahams, Goldstein, Woolley, Mclaughlin, Snowden, Mioshi, Roberts-South, Benatar, HortobáGyi, Rosenfeld, Silani, Ince & Turner, 2017).

Cognitive changes in ALS generally involve executive dysfunction; however, there is evidence that language impairments are common and can occur without executive dysfunction or dementia (Rakowicz, Hodges, 1998). A very recent study reports that in non-demented ALS patients, language impairment occurs in ~23 % of cases, and may be not associated with other motor or non-motor features (Solca, Aiello, Torre, Carelli, Ferrucci, Verde, Ticozzi, Silani, Monti & Poletti, 2023).

The language impairment as extramotor feature of ALS may be present also at early stages and reflect distinct aspects of language processing, specifically in the domains of word naming, syntactic/grammatical processing, and orthographic lexical processing (Pinto-Grau, Donohoe, O'Connor, Murphy, Costello, Heverin, Vajda, Hardiman & Pender, 2021). Syntactic processing impairments are detected in the form of a reduced number of utterances, with shorter sentences (Tsermentseli, Leigh, Taylor, Radunovic, Catani & Goldstein, 2016).

Furthemore, ALS patients show difficulties in connecting one event to the next (local coherence) and in maintaining the topic of the discourse (Ash, Menaged, Olm, McMillan, Boller, Irwin, McCluskey, Elman & Grossman, 2014). Recent evidence from MRI studies confirms that the central nervous system involvement in ALS extends beyond motor areas. Whole brain volume loss and regional frontotemporal atrophy has been reported in ALS patients (Mezzapesa, Ceccarelli, Dicuonzo, Carella, De Caro, Lopez, Samarelli, Livrea & Simone, 2007; Agosta, Chiò, Cosottini, De Stefano & Falini, 2010) (Figura 4). Further, a correlation between extended gray matter involvement, beyond motor areas, and clinical features has been found (Bede, Bokde, Elamin, Byrne, McLaughlin, Jordan, Hampel, Gallagher, Lynch, Fagan, Pender & Hardiman, 2013; Mezzapesa, D'Errico, Tortelli, Distaso, Cortese, Tursi, Federico, Zoccolella, Logroscino, Dicuonzo & Simone, 2013) (Figure 5). This evidence confirms the strict relationships between extra-motor gray matter impairment and cognitive deficit, also including verbal fluency and language deficits, along the ALS-FTD spectrum (Omer, Finegan, Hutchinson, Doherty, Vajda, McLaughlin, Pender, Hardiman & Bede, 2017).

Figure 4 - *Regional cortical Gray matter atrophy in ALS. The main regions with lower GM volume were left middle temporal gyrus and the subgyral region (A), right inferior frontal gyrus (B), and frontal and temporal areas (C). (Mezzapesa et al, AJNR, 2017 n.30)*

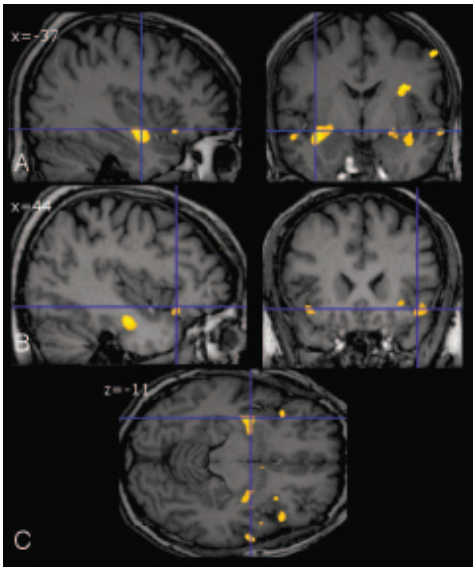
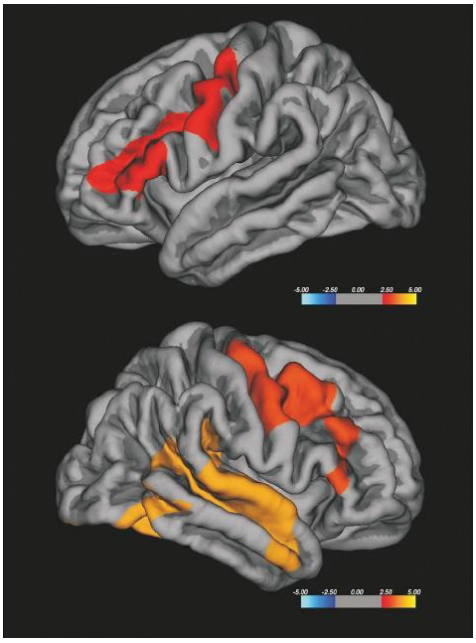


Figura 5 - *Vertex-wise analysis of Cortical Thickness in ALS patients related to healthy controls. Cortical thinning in bilateral precentralcortex, bilateral middle frontal gyrus, right superior temporal and right occipital cortex. The colour bar scale represents t values. (Mezzapesa et al, PlosOne, 2013, n. 31)*



2.5 Multiple Sclerosis

According to the literature, about 40-65 % of the patients with Multiple Sclerosis (MS) develop cognitive deficits, regardless of the duration and severity of the disease (DeLuca, Chiaravalloti & Sandroff, 2020; Benedict, Amato, DeLuca & Geurts, 2020).

Cortical and subcortical brain structures, including the thalamus and cerebellum, play a role in language processing, and the impairment of these regions in MS patients may result in communication disorder. Frequently patients show speech-perceived disorders and dysarthria (Murdoch, Theodoros, 2000). Some patients may have a specific type of aphasia (Demirkiran, Ozeren, Sönmezler & Bozdemir, 2006; Lacour, De Seze, Revenco, Lebrun, Masmoudi, Vidry, Rumbach, Chatel, Verier & Vermersch, 2004). Generally, the language deficits may be of various types, affecting different abilities (verbal fluency, lexical access, language comprehension, etc.) and often associated with other general cognitive deficits such as executive dysfunctions. The most commonly identified symptom is impaired word retrieval in verbal fluency (phonological and semantic) (Portaccio, Amato 2022).

3. *Conclusions*

The extensive group of neurodegenerative diseases share some pathological and clinical signs related to the atrophy of the brain areas connected to specific functions, such as cognitive functions with memory loss, disorientation, language impairment. During the course of neurodegenerative diseases, the insidious deterioration of motor speech, language, and cognition is very frequent, with a significant impact for patients, their caregivers as well for the medical management. In neurodegenerative diseases, language and speech may be impaired for motor dysfunctions affecting vocal respiration, phonation, articulation and prosody, but but also for cognitive-linguisting network failures. Language deficits sometimes may represent a selective and prominent symptom of a neurodegenerative disease, as in Primary Progressive Aphasia (PPA), or may be associated with other cognitive disorders, as in Alzheimer's disease (AD), or with motor-cognitive dysfunction, as in movement's disorders or amyotrophic lateral sclerosis. A great contribution both on the knowledge of neurodegenerative diseases and language networks has recently been provided by advanced MRI techniques.

References

Agosta, F., Chiò, A., Cosottini, M., De Stefano, N., Falini, A., Mascalchi, M., Rocca, M.A., Silani, V., Tedeschi, G. & Filippi M. (2010). The present and the future of neuroimaging in amyotrophic lateral sclerosis. In *American Journal of Neuroradiology*, 31, 1769–1777. https://doi:10.3174/ajnr.A2043

Ash, S., Menaged, A., Olm, C., McMillan, C.T., Boller, A., Irwin D.J., McCluskey, L., Elman, L. & Grossman, M. (2014). Narrative discourse deficits in amyotrophic lateral sclerosis. In *Neurology*, 5, 83(6), 520-8. https://doi:10.1212/WNL.0000000000000670

Baboo, G.K., Prasad, R., Mahajan, P. & Baths, V. (2022). Tracking the Progression and Influence of Beta-Amyloid Plaques Using Percolation Centrality and Collective Influence Algorithm: A Study Using PET Images. In *Annals of Neurosciences*, 29(4), 209-224. https://doi:10.1177/09727531221117633

Baillieux, H., De Smet, H., Paquier, P., DeDeyn, P. & Marien, P. (2008). Cerebellar neurocognition: Insights into the bottom of the brain. In *Clinical Neurology and Neurosurgery*, 110, 763–773. https://doi:10.1016/j.clineuro.2008.05.013

Bang, J., Spina, S. & Miller B.L. (2015). Frontotemporal dementia. In *Lancet*, 24, 386(10004), 1672-82. https://doi:10.1016/S0140-6736(15)00461-4

Bayles, K.A., Tomoeda, C.K., Cruz, R.F. & Mahendra N. (2000). Communication abilities of individuals with late-stage Alzheimer disease. In *Alzheimer Disease and Associated Disorders*, 14(3), 176–181. https://doi:10.1097/00002093-200007000-00009

Bede, P., Bokde, A., Elamin, M., Byrne, S., McLaughlin, R.L., Jordan, N., Hampel, H., Gallagher, L., Lynch, C., Fagan, A.J., Pender, N. & Hardiman, O. (2013). Grey matter correlates of clinical variables in amyotrophic lateral sclerosis (ALS): a neuroimaging study of ALS motor phenotype heterogeneity and cortical focality. In *Journal of Neurology, Neurosurgery and Psychiatry*, 84(7), 766-73. https://doi:10.1136/jnnp-2012-302674

Benedict, R.H.B., Amato, M.P., DeLuca, J. & Geurts, J.J.G. (2020). Cognitive impairment in multiple sclerosis: Clinical management, MRI, and therapeutic avenues. In *The Lancet Neurology*, 19, 860-871. https://doi:10.1016/S14744422(20)30277-5

Booth, J.R., Wood, L., Lu, D., Houk, J.C. & Bitan, T. (2007). The role of the basal ganglia and cerebellum in language processing In *Brain Research*, 16, 1133(1), 136–144. https://doi:10.1016/j.brainres.2006.11.074

Catricalà, E., Della Rosa, P.A., Plebani, V., Perani, D., Garrard, P. & Cappa, S.F. (2015). Semantic feature degradation and naming performance. Evidence from neurodegenerative disorders. In *Brain Language*, 147, 58–65. https://doi:10.1016/j.bandl.2015.05.007

Ceslis, A, Argall, R., Henderson, R.D., McCombe, P.A. & Robinson, G.A. (2020). The spectrum oflanguage impairments in amyotrophic lateral sclerosis. In *Cortex*, 132, 349-360. https://doi:10.1016/j.cortex.2020.09.003

Damier, P., Hirsch, E.C., Agid, Y. & Graybiel, A.M. (1999). The substantia nigra of the human brain II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. In *Brain*, 122(8), 1437–1448. https://doi:10.1093/brain/122.8.1437

DeLuca, J., Chiaravalloti, N.D. & Sandroff, B.M. (2020). Treatment and management of cognitive dysfunction in patients with multiple sclerosis. In *Nature Reviews Neurology*,16, 319–332. https://doi:10.1038/s41582-020-0355-1

Demirkiran, M., Ozeren, A., Sönmezler, A. & Bozdemir, H. (2006). Crossed aphasia in multiple sclerosis. In *Multiple Sclerosis*, 12(1), 116-119. https://doi:org/10.1191/135248506ms1255cr

Ferris, S.H., Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. In *Clinical Interventions in Aging*, 8, 1007-14. https://doi:10.2147/CIA.S39959

Forbes-McKay, K.E., Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. In *Neurological Sciences*, 26, 243–254. https://doi:10.1007/s10072-005-0467-9

Goberman, A.M., Coelho, C. (2002). Acoustic analysis of Parkinsonian speech I: speech characteristics and L-Dopa therapy. In *NeuroRehabilitation*, 17, 237–246, https://doi:10.3233/NRE-2002-17310

Goldstein, L.H., Abrahams S. (2013). Changes in cognition and behaviour in amyotrophic lateral sclerosis: Nature of impairment and implications for assessment. In *The Lancet Neurology*, 12(4), 368–380.https://doi:10.1016/S1474-4422(13)70026-7

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M. & Grossman, M. (2011) Classification of primary progressive aphasia and its variants. In *Neurology*, 15, 76(11), 1006-14. https://doi:10.1212/WNL.0b013e31821103e6

Grossman, M., Crino, P., Reivich, M., Stern, M.B. & Hurtig, H.I. (1992). Attention and sentence processing deficits in Parkinson's disease: The role of anterior cingulate cortex. In *Cerebral Cortex*, 2(6), 513–525. https://doi:10.1093/cercor/2.6.513

Grossman, M., Kalmanson, J., Bernhardt, N., Morris, J., Stern, M.B., Hurtig, H.I. (2000). Cognitive resource limitations during sentence comprehension in Parkinson's disease. In *Brain Language*,1, 73(1), 1-16. https://doi:10.1006/brln.2000.2290

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. In *Frontiers in Psychology*, 12, 4, 416. https//doi:10.3389/fpsyg.2013.00416

Hertrich, I., Dietrich, S. & Ackermann, H. (2020). The Margins of the Language Network in the Brain. In *Frontiers in Communication*, 5, 519955. https://doi.org/10.3389/fcomm.2020.519955

Highnam, C.L., Bleile, K.M. (2011). Language in the cerebellum. In *American Journal of Speech-Language Pathology*, 20(4), 337-347. https://10.1044/1058-0360(2011/10-0096)

Hou, Y., Dan, X., Babbar, M., Wei, Y., Hasselbalch, S.G., Croteau, D.L. & Bohr, V.A. (2019). Ageing as a risk factor for neurodegenerative disease. In *Nature Reviews Neurology*,15(10), 565-581. https://doi:10.1038/s41582-019-0244-7

Kuhnlein, P., Gdynia, H., Sperfeld, A., Lindner-Pfleghar, B., Ludolph, A.C., Prosiegel, M. & Riecke, A. (2008). Diagnosis and treatment of bulbar symptoms in amyotrophic lateral sclerosis. In *Nature ClinicalPractice Neurology*, 4(7), 366–374. https://doi:10.1038/ncpneuro0853

Lacour, A., De Seze, J., Revenco, E., Lebrun, C., Masmoudi, K., Vidry, E., Rumbach, L., Chatel, M., Verier, A. & Vermersch, P. (2004). Acute aphasia in multiple sclerosis: A multicenter study of 22 patients. In *Neurology*, 23, 62(6), 974-7. https://doi:10.1212/01.wnl.0000115169.23421.5d

Lomen-Hoerth, C., Murphy, J., Langmore, S., Kramer, J.H., Olney, R.K. & Miller, B. (2003). Are amyotrophic lateral sclerosis patients cognitively normal? In *Neurology*, 60(7), 1094–1097. https://doi:10.1212/01.wnl.0000055861.95202.8d

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D. & Stadlan, E.M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. In *Neurology*, 34:939-944. https://doi:10.1212/wnl.34.7.93.9

Mezzapesa, D.M., Ceccarelli, A., Dicuonzo F., Carella, A., De Caro, M.F., Lopez, M., Samarelli, V., Livrea, P. & Simone I.L. (2007). Whole-brain and regional brain atrophy in amyotrophic lateral sclerosis. In American Journal of Neuroradiology, 28(2):255-9. PMID: 17296989; PMCID: PMC7977419

Mezzapesa, D.M., D'Errico, E., Tortelli, R., Distaso, E., Cortese, R., Tursi, M., Federico, F., Zoccolella, S., Logroscino, G., Dicuonzo, F. & Simone, I.L. (2013). Cortical thinning and clinical heterogeneity in amyotrophic lateral sclerosis. In *PLoSOne*, 20, 8(11), e80748. https://doi:10.1371/journal.pone.0080748

Mitchell, J.D., Borasio, G.D. (2007) Amyotrophic lateral sclerosis. In *The Lancet*, 369, 2031– 2041. https//doi:10.1016/S0140-6736(07)60944-1

Murdoch B.E., Theodoros D.G. (2000). Speech and language disorders in multiple sclerosis. In Wiley, J. & Sons Ltd (a cura di), *Speech and Language Disorders in Multiple Sclerosis*,. ISBN 10: 1861561008 / ISBN 13: 9781861561008

Nadeau, S.E., Crosson, B. (1997). Subcortical aphasia. In *Brain Language*, 58(3):355–40227. https://doi:10.1006/brln.1997.1707

Nadeau, S.E. (2021). Basal Ganglia and Thalamic Contributions to Language Function: Insights from A Parallel Distributed Processing Perspective. In *Neuropsychology Review* 31(3), 495-515. https://doi:10.1007/s11065-020-09466-0

Neary, D., Snowden, J.S. & Mann, D.M.A. (2000). Cognitive change in motor neurone disease/amyotrophic lateral sclerosis (MND/ALS). In *Journal of the Neurological Sciences*, 180(1), 15–20. https://doi:10.1016/s0022-510x(00)00425-1

Omer, T., Finegan, E., Hutchinson, S., Doherty, M., Vajda, A., McLaughlin, R.L., Pender, N., Hardiman, O. & Bede, P. (2017). Neuroimaging patterns along the ALS-FTD spectrum: a multiparametric imaging study. In *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 18(7-8), 611-623. https://doi:10.1080/21678421.2017.1332077

Peng, C., Trojanowski, J.Q. & Lee, V.M. (2020). Protein trnsmission in neurodegnerative diseases. In *Nature Reviews Neurology*, 16(4), 199-212, https://doi:10.1038/s41582-020-0333-7

Pinto-Grau, M., Donohoe, B., O'Connor, S., Murphy, L., Costello, E., Heverin, M., Vajda, A., Hardiman, O. & Pender, N. (2021). Patterns of Language Impairment in Early Amyotrophic Lateral Sclerosis. In *Neurology Clinical Practice*, 11(5), e634-e644. https://doi:10.1212/CPJ.0000000000001006

Portaccio, E., Amato, M.P. (2022). Cognitive Impairment in Multiple Sclerosis: An Update on Assessment and Management. In *NeuroSci*, *3*(4), 667-676. https://doi.org/10.3390/neurosci 3040048

Rakowicz, W.P., Hodges, J.R. (1998). Dementia and aphasia in motor neuron disease: An underrecognised association? In *Journal of Neurology*, Neurosurgery and Psychiatry, 65(6), 881e889. https://doi:10.1136/jnnp.65.6.881

Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J., van Swieten, J., Seelaar, H., Dopper, E.G., Onyike, C.U., Hillis, A.E., Josephs, K.A., Boeve, B.F., Kertesz, A., Seeley, W.W., Rankin, K.P., Johnson, J.K., Gorno-Tempini, M.L., Rosen, H., *et al.* (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. In *Brain*, 134(Pt 9), 2456–2477. https://doi:10.1093/brain/awr179

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H. & Obeso, J.A. (2010). Goal-directed and habitual control in the basal ganglia: Implications for Parkinson's disease. In *Nature Reviews Neuroscience*, 11(11), 760–772. https://doi:10.1038/nrn2915

Ringholz, G.M., Appel, S.H., Bradshaw, M., Cooke, N.A., Mosnik, D.M. & Schulz, P.E. (2005). Prevalence and patterns of cognitive impairment in sporadic ALS. In *Neurology*, 65(4), 586e590. https://doi.org/10.1212/01.wnl.0000172911.39167.b6

Solca, F., Aiello, E.N., Torre, S., Carelli, L., Ferrucci, R., Verde, F., Ticozzi, N., Silani, V., Monti, A. & Poletti, B. (2023). Prevalence and determinants of language impairment in non-demented amyotrophic lateral sclerosis patients. In *European Journal of Neurology*, 30(3), 606-611. https://doi:10.1111/ene.1565

Strong, M.J., Abrahams, S., Goldstein, L.H., Woolley, S., Mclaughlin, P., Snowden, J. Mioshi, E., Roberts-South, A., Benatar, M., HortobáGyi, T., Rosenfeld, J, Silani, V., Ince, P.G. & Turner, M.R. (2017). Amyotrophic lateral sclerosis – frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. In *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 18(3-4), 153-174. https://doi:10.1080/21678421.2016.1267768

Taler, V., Phillips, N.A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. In *Journal of Clinical and Experimental Neuropsychology*, 30, 501–556. https://doi:10.1080/13803390701550128

Taylor, L.J., Brown, R.G., Tsermentseli, S., Al-Chalabi, A., Shaw, C.E., Ellis, C.M., Leigh, P.N. & Goldstein, L.H. (2013). Is language impairment more common than executive dysfunction in amyotrophic lateral sclerosis? In *Journal of Neurology, Neurosurgery and Psychiatry*, 84(5), 494-8. https://doi:10.1136/jnnp-2012-303526

Tsermentseli, S., Leigh, P.N., Taylor, L.J., Radunovic, A., Catani, M. & Goldstein, L.H. (2016). Syntactic processing as a marker for cognitive impairment in amyotrophic lateral sclerosis. In *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 1, 69–76. https://doi:10.3109/21678421.2015.1071397

Vigo, I., Coelho, L. & Reis, S. (2022) Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review. In *Bioengineering (Basel)*,11, 9(1), 27. https://doi:10.3390/bioengineering9010027

Whitwell, J.L. (2019). FTD spectrum: Neuroimaging across the FTD spectrum. In *Progress in Molecular Biology and Translational Science*, 165, 187–223. https://doi:10.1016/bs.pmbts.2019.05.00

DOI: 10.17469/O2111AISV000003

SERGE PINTO

Speech motor control from its impairment: the relevance of studying dysarthria in Parkinson's disease

Dysarthria is a collective name for a group of neurologic motor speech disorders, resulting from central and/or peripheral nervous system abnormalities. Hypokinetic dysarthria in Parkinson's disease presents with prosodic insufficiency, related to a monotony of pitch and intensity, a reduction of accentuation, variable speech rate and possible phoneme imprecision. In most cases, the voice is harsh and breathy. Many patients complain about speech impairments, which affect their communication in daily living activities. There is still a need for further research of dysarthria in Parkinson's disease, for a better understanding of its pathophysiology, treatment response, and relationship with other functions/symptoms. The aim of this paper is to provide an insight into the investigation of dysarthria in Parkinson's disease from two complementary perspectives: from a clinical perspective, to help design patient individualised management; from a neurolinguistics point of view, to help better understand the relative contribution of basal ganglia in speech motor control.

Keywords: Parkinson's disease, dysarthria, speech, treatment, challenges.

1. *Introduction*

Speech motor control is an important part of communication, and the breakdown of such motor control can result in speech impairment. Motor speech disorders refer to a set of signs affecting the control and production of speech consequent to neurological impairment underlying movement disorders. Therefore, they can be distinguished from higher-order language difficulties (i.e., aphasias) and characterized by an approach that dichotomizes them into two entities: apraxia of speech and dysarthria.

Authors regularly emphasize historical elements, specifying that the term "dysarthria" does not only refer to "arthritic" or supra-laryngeal articulatory disorders (e.g. Auzou, 2007; Duffy, 2007). This was the case for many years; dysarthria was then used to describe a primary disorder of speech articulation (Brain, 1965). Since the early 20th century, several authors (cited in Duffy, 2013) grouped the consequences of lesions of the sensorimotor speech system under the terms either dysarthria, i.e. articulation disorders (with the prefix *-dys* for dysfunction and the Latin root *arthri* for articulation), or anarthria, i.e. absence of articulation, the prefix *-a* being privative (Liepmann, 1913; Head, 1926). This Anglo-Saxon terminology of speech disorders, based on the stance of Darley, Brown & Aronson (1975), has been mainstream since the second half of the 20th century.

Dysarthria is "a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system. It designates problems in oral communication due to paralysis, weakness, or incoordination of the speech musculature" (Darley, Brown & Aronson, 1969a: 246). Thus, dysarthria refers to "a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory, and prosodic aspects of speech production" (Duffy, 2005: 3). This is a current definition of dysarthria, confined to a neurological origin of the disorder, specific, and excluding mechanically induced impairments.

The multiple and often different uses of the term dysarthria can be confusing, and it remains important to remember that dysarthria is a sign that is part of a broader clinical picture, related to the overall symptomatology of a specific movement disorder, associated with a lesion of the central or peripheral nervous system. Clinically, the concept of arthritic disorder encompasses all supra-laryngeal articulatory disorders, in both adults and children, of neurological (e.g., neurodegenerative disease) or non-neurological origin (e.g., ablative surgery of the tongue). Dysphonic disorders, on the other hand, include all forms of laryngeal dysfunction, regardless of their origin. This dissociation between conceptual and clinical terminology would require interdisciplinary consensus to clarify certain aspects of this terminology; in fact, the term "neurodysarthrophony" frequently appears to account for both articulatory and phonatory disorders related to Parkinsonian dysarthria (Ackermann & Ziegler, 1989; Klostermann, Ehlen, Vesper, Nubel, Gross, Marzinzik, Curio & Sappok, 2008), or even the term "neuroarthropneumophony" (Moreau, Pennel-Ployart, Pinto, Plachez, Annic, Viallet, Destée & Defebvre, 2011), used since 1957 but rarely used since (Grewel, 1957). The term "neurodysarthrophony" might be the one that most precisely defines the current concept of dysarthria, i.e., an arthritic (articulatory) and/or pneumo-phonatory dysfunction of neurological origin. The terminological orthogonality between dysarthria/dysphonia can be problematic because, while the term dysphonia is identified as a voice disorder, dysarthria is often mistakenly equated with an exclusive arthritic disorder, an erroneous analogy since articulatory disorders resulting, for example, from ablative orofacial surgery do not fall within the scope of dysarthria (Pinto & Ghio, 2008).

Among other possibilities, two approaches to examining dysarthria in movement disorders could be considered (Pinto, Chan, Guimarães, Rothe-Neves & Sadat, 2017). On the one hand, one can adopt a disease-based approach, which implies that pathophysiological processes are at the origin of the motor signs that contribute, maybe exclusively, to speech alterations; On the other hand, a neurolinguistics-based approach considers that motor speech disorders are the result of alterations dependent on modifications of linguistic processes that have emerged along the progression of the disease. From the former language-independent perspective (i.e., disease-based), dysarthria needs to be assessed, to be managed. From the latter language-specific perspective (i.e., neurolinguistics-based), dysarthria is a model

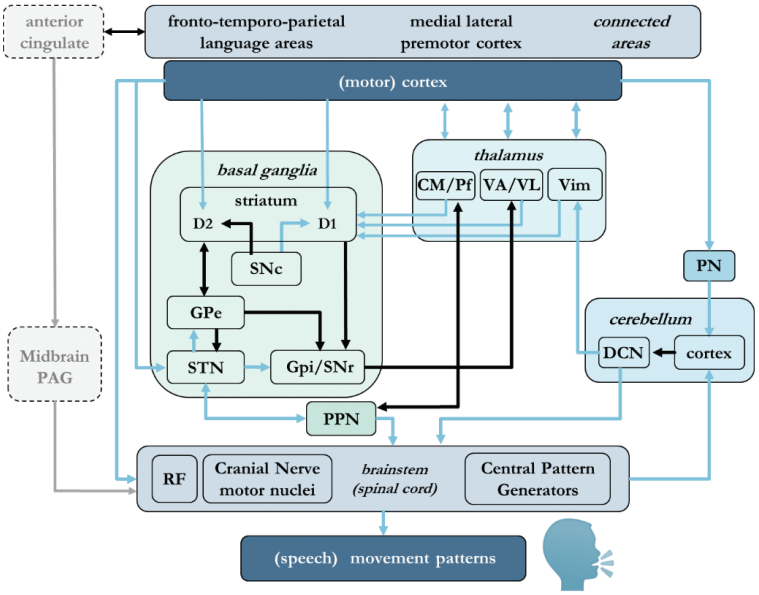
of pathological state which study could help refining neurolinguistics models of speech production. Biasing towards one of the two approaches only would narrow the impact of the findings, and it is reasonable to argue that both approaches are complementary and much needed to provide the most thorough description and analysis of dysarthria understanding.

2. *From speech motor control to motor speech disorders*

As a very subjective starting point, we could remember that "speech motor control refers to the systems and strategies that regulate the production of speech, including the planning and preparation of movements (sometimes called motor programming) and the execution of movement plans to result in muscle contractions and structural displacements" (Kent, 2000). A thorough neuroimaging review synthesis provided by Price (2012) summarized from a historical perspective the brain areas showed to be involved in spoken and written language, as well as auditory processes. As regards speech production, named in the review overt articulation, she highlighted cortical regions such as the anterior cingulate cortex, the supplementary motor area and motor/sensorimotor cortices. She also mentioned that the cerebellum, the insula and subcortical areas, such as the putamen and the thalamus, were involved in the timing and control of motor activity. This statement was quite important since mainstream psycholinguistic models, at that time, mainly focused on cortical areas to the detriment of subcortical ones (e.g., Hickock & Poeppel, 2004; Poeppel, Emmorey, Hickok & Pylkkänen, 2012).

A more recent review spotted similar findings, though presented differently (Hertrich, Dietrich & Ackermann, 2020). Notably, the authors provided an illustrative sketch that links brain areas among a "language network" and implicated in several linguistic processes: auditory, motor, lexical, semantic, syntactic, and aspects related to broader social cognition. Again here, the authors highlighted the importance of subcortical areas such as the basal ganglia, the thalamus and the cerebellum. Focusing even more on subcortical contributions to motor speech, Ziegler & Ackermann (2017) summarized nicely the concept in a motor, neurophysiological model that integrates, in addition to an innate limbic vocal system (involving the anterior cingulate and the midbrain periaqueductal grey matter), the two classic motor regulatory loops (the cortico-striatal and cortico-cerebellar ones), which modulate the pyramidal motor tract. Such a model can be implemented in another one (Wichmann, 2019), which is more traditionally considered for the understanding of movement disorders like parkinsonism (Figure 1).

Figure 1 - *A neurophysiological/neurolinguistics view of speech production. Adapted by combining figures from Ackermann (2008), Ziegler & Ackermann (2017) and Wichmann (2019). CM/Pf: CentroMedian-ParaFascicular nuclei complex; DCN: Deep Cerebellar Nuclei; D1/D2: striatal dopamine D1 and D2 receptors; GPe: external Globus Pallidus; GPi: internal Globus Pallidus; PAG: PeriAqueductal Gray; PN: Pontine Nucleus; PPN: PedunculoPontine Nucleus; RF: Reticular Formation; SNc: Substantia Nigra pars compacta; SNr: Substantia Nigra pars reticulata; STN: SubThalamic Nucleus; VA: Ventral Anterior nucleus; VL: Ventral Lateral nucleus; Vim: Ventral intermediate nucleus*



Connecting speech motor control to motor speech disorders would mean trying to complement neurophysiological models by integrating specificities – and dysfunctions – related to speech motor control, or vice-versa, to refine neurolinguistic models by integrating neurophysiological motor control circuitries – and their dysfunctions. It is the objective of many recent neurolinguistics or computational models, such as the ACTion-based model of speech production, speech perception, and speech acquisition (ACT) (Kröger *et al.*, 2012) or the Directions Into Velocities of Articulators (DIVA) model (Guenther, 1994; Golfinopoulos, Tourville & Guenther, 2010; Tourville & Guenther, 2011). Furthermore, as reminded by Hertrich *et al.* (2020), the role of subcortical circuits for speech motor control is known principally from clinical studies on dysarthrias after basal ganglia or cerebellar disorders. There is indeed a very tight connection and a long history between the elaboration of fundamental models of speech motor control and the report of clinical cases.

Different forms of motor speech disorders correspond to different aetiologies and pathophysiologies. They are part of a broader range of motor deficits inherent to a specific clinical picture: the idea is, therefore, to match the specific combination of signs defining the observed motor speech disorder to a precise topology of brain

lesion. Several attempts to classify dysarthrias have been initiated, both benefiting from and encountering challenges in multi- and interdisciplinary approaches: it is indeed difficult to synthesize neuroanatomical, etiological, semiological, and perceptual data into a simple and exhaustive classification. However, it has been acknowledged that the perceptual characteristics of each dysarthria depended more on the location of the neurological lesion, the anatomical substrate of the associated pathology than on its aetiology (Luchsinger & Arnold, 1965).

3. *Clinical-perceptual classification of dysarthrias by Darley and colleagues*

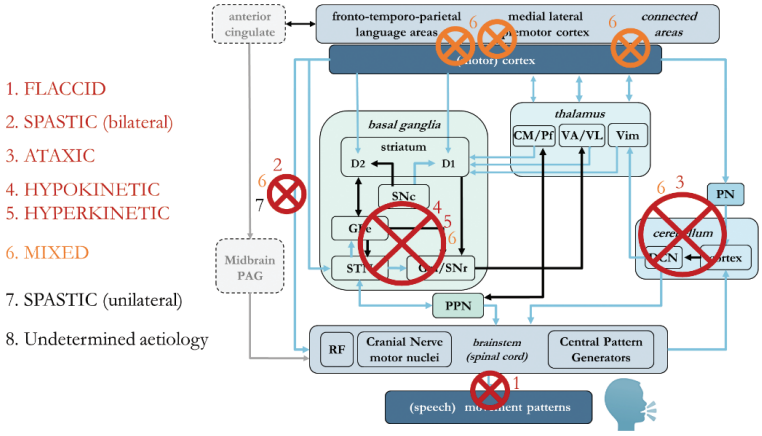
There is no classification of dysarthrias based solely on their semiotic description. The one originating from the Mayo Clinic in Cleveland (USA) remains the most widely recognized and used (Darley *et al.*, 1969a; 1969b; 1975). It offers a pragmatic description based on the grouping of perceptually evaluated anomalies and still represents a consensual, easy-to-understand and practical way to describe speech impairment in movement disorders. The description of six types of dysarthrias laid the groundwork for pathophysiological hypotheses aiming to associate a characteristic speech motor disorder with a precise anatomical lesion (Figure 2). Two additional dysarthria types have been added by Duffy (2005) to this classification: dysarthria resulting from unilateral upper motor neuron disorder (while spastic dysarthria is consequent to bilateral dysfunction), and dysarthria from undetermined aetiology.

3.1 Classification procedure

The classification was made from the study of 212 patients distributed into seven pathological groups, each of them meant to represent a typical pathophysiology, associated with one of the six dysarthria types (Figure 2):

- flaccid dysarthria for bulbar palsy (dysfunction of the motoneuron);
- spastic dysarthria for pseudo-bulbar palsy (dysfunction of the upper motoneuron);
- ataxic dysarthria for cerebellum dysfunctions;
- hypokinetic dysarthria for parkinsonism (basal ganglia dysfunction);
- hyperkinetic dysarthria for both chorea and dystonia (basal ganglia dysfunctions);
- mixed dysarthria for amyotrophic lateral sclerosis, ALS (diffuse brain dysfunctions).

Figure 2 - *A pathophysiological/neurolinguistic view of dysarthrias. Localisations of cerebral lesions/dysfunctions related to the different pathological groups have been identified in the normal functioning model proposed in Figure 1. Such localisations are identified in coloured circles, and associated to dysarthria types*



Among the 212 patients studied, 30 were included in each group, except for the Parkinsonian group which comprised 32 patients. Patients were required to read a short text, the Grandfather passage, and speech productions were recorded. The evaluation consisted of the three authors of the classification (Darley, Brown, and Aronson) paying attention to the characterization of 38 dimensions covering seven speech features: pitch (n=4 dimensions), loudness (n=5), voice quality (n=9), respiration (n=3), prosody (n=10), articulation (n=5), and an "overall" feature (n=2) that provided a general impression of the intelligibility and the "bizarreness" of speech (Table 1).

Table 1 - *Speech and voice features perceptually evaluated by Darley, Brown and Aronson for the classification of dysarthrias*

Pitch	
1	Pitch level
2	Pitch breaks
3	Monopitch
4	Voice tremor

Loudness	
5	Monoloudness
6	Excess loudness variation
7	Loudness decay
8	Alternating loudness
9	Loudness (overall)

"Overall impression"	
37	Intelligibility
38	Bizareness

Vocal quality	
10	Harsh voice
11	Hoarse (wet) voice
12	Breathy voice (continuous)
13	Breathy voice (transient)
14	Strained-strangled voice
15	Voice stoppages
16	Hypernasality
17	Hyponasality
18	Nasal emission

Respiration	
19	Forced inspiration-expiration
20	Audible inspiration
21	Grunt at end of expiration

Prosody	
22	Rate
23	Phrases short
24	Increase of rate (in segments)
25	Increase of rate (overall)
26	Reduced stress
27	Variable rate
28	Intervals prolonged
29	Inappropriate silences
30	Short rushes of speech
31	Excess and equal stress

Articulation	
32	Imprecise consonants
33	Phonemes prolonged
34	Phonemes repeated
35	Irregular articulatory breakdown
36	Vowels distorted

Some criteria are not discriminating at all: a hoarse voice, monotony, and imprecision of consonants describe most types of dysarthria. Conversely, excessive intensity variations are typical of dystonic voice, as are vocal arrests; Parkinsonian speech, on the other hand, exhibits very characteristic paroxysmal accelerations. By correlating certain deviant criteria referring to the same physiological function, eight syndromic clusters were described, which characterise the pathological groups considered in the study and their associated dysarthrias (Table 2). One may question why, in Parkinson's disease, for example, only the syndromic cluster of prosodic incompetence has been highlighted. This classification certainly shows limitations (see below § 3.3).

3.2 Brief description of dysarthrias

Flaccid Dysarthria is characteristic of the bulbar pathological group. It results from damage to the final common motor pathway, which includes peripheral nerves (arising from the corticobulbar and/or corticospinal tracts), the neuromuscular junction (e.g., as in myasthenia), and/or the speech effector muscles (e.g., as in myotonic dystrophy of Steinert, oculopharyngeal myopathy, Duchenne's dystrophy). It is characterised by low tone, paralysis, or hypotonia.

Spastic Dysarthria is the pseudobulbar one, resulting from bilateral damage to the first neuron connecting the motor cortex to the brainstem relay. The corticobulbar tract is thus bilaterally affected, with the conditions mainly related to lacunar states (small strokes). The term "spastic" in this context refers to an exaggerated and permanent increase in resting muscle tone. Pseudobulbar syndromes appearing in other pathologies (e.g., ALS, multiple sclerosis [MS], progressive supranuclear palsy [PSP]) and involving impairment of speech production from more diffuse lesions lead to mixed dysarthria (cf. below).

Ataxic Dysarthria results from damage to the cerebellum and/or cerebellar pathways. Lesions can be of various natures (degenerative, vascular, traumatic, neoplastic, inflammatory, toxic, or metabolic). The lack of coordination of voluntary movements, characteristic of cerebellar deficit, is expressed notably in articulatory imprecision and prosodic excesses.

Hypokinetic Dysarthria results from degenerative damage to a part of the basal ganglia, causing a more global dysfunction of the motor regulation loop involving the basal ganglia. Akinesia (inability to move or slow initiation of movement), bradykinesia (slow execution of movement), and hypokinesia (movements of low amplitude) characterize the pathological group centred around Parkinson's disease (PD). Dysarthrias of other parkinsonian syndromes (multiple system atrophy [MSA], PSP, Lewy body dementia [LBD], or even corticobasal degeneration [CBD]) share some hypokinetic characteristics with other types of dysarthria (hyperkinetic, spastic, etc.), often taking on a mixed character.

Hyperkinetic Dysarthria is characteristic of the dystonia and choreoathetosis pathological groups. Dystonia, properly speaking, refers to rather slow movements and twists leading to abnormal postures. Dystonia can be focal (laryngeal, oro-mandibular, cervical), more extensive (generalized or segmental dystonia), primary,

or secondary (resulting from neuroleptic intake, perinatal anoxia, post-traumatic, post-stroke, metabolic). Abnormal choreiform movements are involuntary, rapid, and non-stereotyped. Their aetiology can be degenerative, as in the case of Huntington's disease, toxic, or post-infectious. Hyperkinetic dysarthria also results from dysfunction of the basal ganglia, involving, as the name suggests, the occurrence of large, involuntary, and uncontrollable movements. Spasmodic dysphonia, often associated with essential tremor, is also an example of hyperkinetic dysarthria.

In Darley, Brown, and Aronson's classification, the pathological group that illustrated mixed dysarthria consisted of ALS patients. Mixed dysarthrias combine speech disorders resulting from the involvement of several neurological systems; mixed dysarthric speech is characteristic of pathological groups such as ALS, MS, traumatic brain injuries, but also atypical parkinsonian syndromes (e.g., LBD, MSA, PSP).

Later, two additional categories complemented this classification (Duffy, 2005: 13-14): dysarthrias due to unilateral damage to the first motor neuron, allowing for a distinction concerning a specific hemisphere compared to spastic dysarthria involving bilateral damage; and dysarthrias of undetermined aetiology.

Table 2 - *Syndromic clusters described by Darley, Brown et Aronson (according to Auzou, 2007). AD: Ataxic Dysarthria; ALS: Amyotrophic Lateral Sclerosis; Bulb.: Bulbar; Cereb.: Cerebellar; Park.: Parkinsonism; Chor.: Chorea; Dyst.: Dystonia; FD: Flaccid Dysarthria; HoD: Hypokinetic Dysarthria; HyD: Hyperkinetic Dysarthria; Ps-bulb.: Pseudobulbar; SD: Spastic Dysarthria*

Pathological group	Bulb.	Ps-bulb.	Cereb.	Park.	Chor.	Dyst.	ALS
Dysarthria type	FD	SD	AD	HoD	HyD		Mixed
Articulatory imprecision							
Prosodic excess							
Prosodic insufficiency							
Articulatory and resonatory incompetence							
Phonatory stenosis							
Phonatory incompetence							
Resonatory incompetence							
Phonatory and prosodic insufficiency							

3.3 Limits of such dysarthria classification

The classification by Darley aims to be exhaustive but unfortunately does not allow the systematic placement of the observed disorder into a predefined category: some evaluation criteria are non-discriminatory, some pathological expressions take different forms depending on the patients, and the aetiologies are diverse for the same pathology. For example, hypokinetic dysarthria is expected to characterise PD speech, but it has been described, without any precision, from patients suffering from

"parkinsonism". Oppositely, mixed dysarthria has been described while evaluating patients with ALS only, whereas other affections such as atypical parkinsonism are also examples of mixed dysarthria.

A critical view of this classification of dysarthrias was proposed by Auzou (2007), emphasizing its undeniable contribution to the detailed description of dysarthrias but also highlighting the limitations of this classification: difficulty in the differential diagnosis of dysarthrias based on deviant criteria, lack of duplication work, approximate methodologies for constituting pathological groups and clusters, listening conditions that were not blinded to the diagnoses. Nevertheless, even though the results of this perceptual analysis are difficult to reproduce, this evaluation method is still considered the method of choice (Duffy, 2005; 2013). All these considerations pave the way for increasingly numerous studies on dysarthria, whose description and classification are still to be perfected. Over time and despite limitations, including the almost absence of reproducibility, this description by Darley and colleagues remains highly favoured in clinical practice (Kent, Kent, Weismer & Duffy, 2000), at least concerning hypokinetic dysarthria. Auditory-perceptual evaluation remains an important and appropriate clinical tool, as it allows for the description of the multiple dimensions of dysarthric speech: a congruence study among different listening judges using Mayo Clinic criteria strengthens this argument (Bunton, Kent, Duffy, Rosenbek & Kent, 2007).

Certainly, the clinical-perceptive description and classification from the Mayo Clinic have been contextually important from a diagnostic perspective; however, concerning the aetiology and management of these disorders, studying and assessing physiological dysfunctions at each level of production (breathing, phonation, resonance, articulation, prosody) remains crucial for the physiological understanding of dysarthrias. Indeed, the instrumental approach to studying dysarthrias is essential, readily focusing on exploring movement disorders of speech organs: aerodynamic, electromyographic, cineradiographic, kinematic, acoustic measurements, functional brain imaging, or even involving modelling may be useful (Duffy, 2007: 18).

4. *Features of dysarthric speech in Parkinson's disease*

In the seminal description of the disease made by Parkinson (1817), who described thoroughly six patients aged between 50 and 72 years old, three of them presented with speech deficits. The words of James Parkinson were the following: "His words are now scarcely intelligible", referring to a marked loss of speech intelligibility; "The power of articulation is lost", suggesting a disorder of supra-laryngeal articulation; and "Speech was very much interrupted" reporting a deficit of fluency and alteration of prosody in its temporal aspects.

4.1 The timeline of dysarthria in Parkinson's disease

When speech impairment in PD was reported by James Parkinson, no reference to the timing of appearance was made. It was only mentioned that such impairment occurs "as time and the disease proceed" (Parkinson, 1817: 5). Selby (1968) proposed a caudorostral progression of dysarthria, from the respiratory system to the larynx and the supralaryngeal articulation. Selby reported that "*The initial defect in the untreated patient is a failure to control respiration for the purpose of speech and there follows a forward progression of articulatory symptoms involving larynx, pharynx, tongue and finally lips*". The speech deficits in PD seem to start early in the disease timeline with hypophonia and worsen with time to include dysphonia, dysprosody and supralaryngeal dysarthria. From a functional point of view, speech changes in PD spread from loudness to voice and intelligibility. So dysarthria is not a late symptom in PD, but rather a behavioural marker.

The timeline for dysarthria may be tightly connected to the one proposed by Braak with the stage progression of PD (Braak, Del Tredici, Rüb, de Vos, Jansen Steur & Braak, 2003). Particularly, one may find here some pathophysiological origin of the first stage of dysarthria in PD, so namely hypophonia, with the reported early dysfunction of the vagus nerve. Indeed, vagal innervation is crucial for the diaphragm, the core muscle involved in respiration, and by extension to the respiratory control of speech and speech loudness.

4.2 Characteristics of speech in PD: a summary

Table 3 summarizes the ten deviant dimensions identified perceptually by Darley *et al.* (1969a; 1969b) that characterised the most speech in PD patients. Hypokinetic dysarthria involves laryngeal deficits, the voice being harsh, often with a low pitch. Voice breaks, including tremor, can occur, and voice can be also breathy, mainly since vocal folds are quite rigid and patients have difficulties in closing the vocal tract when voicing, implying consequently an air leak that can be heard during speech. A feature that has not been pointed out by Darley and colleagues is the hypophonic nature of PD speech (Canter, 1963). Imprecise articulation contributes to the lack of intelligibility, together with the deficit in respiratory control. From a more suprasegmental perspective, patients have a monotonous voice, use reduced stress and produce short rushes of speech. In other words, speech temporal organisation is disturbed, and speech festination, freezing, or even pseudo-stuttering can also be observed. Several recent reviews can be consulted for further details (e.g., Dashtipour, Tafreshi, Lee & Crawley, 2018; Brabenec, Mekyska, Galaz & Rektorova, 2017; Moreau & Pinto, 2019).

Table 3 - *Deviant dimensions of speech in Parkinson's disease patients, identified perceptually by Darley, Aronson & Brown (1969a; 1969b). Alterations impact prosodic dimensions (in blue), phonatory charateristics (in dark red), temporal organization (in grey) and supralaryngeal articulation deficits (in orange)*

Rank	Dimension
1	Monopitch
2	Reduced stress
3	Monoloudness
4	Imprecise consonants
5	Inappropriate silences
6	Short rushes
7	Harsh voice
8	Breathy voice (continuous)
9	Low pitch
10	Variable rate

4.3 The challenge of acoustic automatic analyses of dysarthric speech

Studying dysarthria might start often by identifying the communication deficits reported by the patients, and the consequences of speech impairment in their daily living activities. However, this needs to take into account the contributing factors that may exacerbate speech and communication deficits. It is also of utmost importance to measure the speech impairment per se, and for that, several means are available. Among them, a renewed interest has been paid to automated acoustic analyses in recent years. The strength of such analyses is to relate specific acoustic measures, obtained with dedicated speech tasks, to specific dimensions of speech. Therefore, together with dedicated guidelines for speech recording (Rusz, Tykalová, Ramig & Tripoliti, 2021), automated analyses can help to assess and monitor speech changes across PD. This is of particular interest for PD diagnosis and management, and to disentangle acoustic specificities or ambiguities between PD and atypical parkinsonism.

Some specific signs are often reported when referring to atypical parkinsonism (Mulroy, Stamelou & Bathia, 2019), which includes more bilateral signs, that appear earlier in the course of the disease; axial signs (such as gait disorder, postural instability, dysarthria, dysphagia, neuropsychological deficits, etc.) are often more severe than idiopathic PD, and the response to levodopa treatment is usually reduced. Regarding speech impairment, dysarthria is mixed, with additional spastic or ataxic features to the hypokinetic signs that characterise PD. Clinical assessment may refer to 'growling' dysarthria that can become constant groaning in PSP. In the case of MSA, a high-pitched and squeaky voice can be noted. Some striking abnormalities can be noted: Involuntary productions in PSP, and Stridor (MSA). In CBD, orofacial apraxia is very frequent. Also, in atypical parkinsonism, dysarthria is often associated with dysphagia that is early severe when compared to PD, and with

neuropsychological deficits that encompass language difficulties that can also be present in PD, but in later stages and later on in the progression of the disease. One recent study, for example, confirmed the predominance of hypokinetic dimensions in PD dysarthria, ataxic features in MSA and spastic in PSP (Kowalska-Taczanowska, Friedman & Koziorowski, 2020). As the authors stated, acoustic voice analysis is a very sensitive and non-invasive tool that can provide objective information for the assessment of different speech components. Such data, at some point, can be essential for the differential diagnosis of Parkinsonian syndromes. In the same vein, it is possible to go further, trying to distinguish and quantify specific features of MSAp and MSAc for example. Interestingly, in this study by Rusz, Tykalová, Salerno, Bancone, Scarpelli & Pellecchia (2019), audible inspirations observed in MSAc was the only individual speech pattern that allowed the discrimination of parkinsonian and cerebellar variants, despite strong overlaps between both MSA groups.

5. *Effects of treatments on speech in Parkinson's disease*

It is a fact: dysarthria is a symptom difficult to manage in PD (Moreau & Pinto, 2019). Dysarthria is usually considered as a symptom partially responsive to treatments that may, or may not, induce additional speech deficits unrelated to the disease itself, as treatment side effects.

5.1 Levodopa and medication strategies

Medication strategies have mitigated effects; dopaminergic medication may improve some speech dimensions but also contribute to the exacerbation of others. When it comes to medication, it is all about disease and treatment durations: the longer, the worse. Speech response to levodopa is variable, and depends on both disease and treatment durations, as well as the stage of the disease (De Letter, Santens, De Bodt, Van Maele, Van Borsel & Boon, 2007; Skodda, Grönheit, Mancinelli & Schlegel 2013). While levodopa can improve some components of articulatory deficits, it is not the case for overall speech intelligibility (e.g., Brabenec *et al.*, 2017). Considering longitudinal studies, speech performance well-reflected improvement/deterioration of motor manifestations in dependence on the L-dopa doses in the early stages of the disease (Rusz, Tykalová, Klempíř, Čmejla & Růžička, 2016), while continuously declined in later stages with poor response to L-dopa (Skodda, Visser & Schlegel 2010).

5.2 Neurostimulation

Managing dysarthria following deep brain stimulation is often restricted to avoid a worsening of speech disorders and to hope for potential positive effects on speech dimensions, such as respiratory control and/or loudness (Tsuboi, Watanabe, Tanaka, Ohdake, Yoneyama, Hara, Nakamura, Watanabe, Senda, Atsuta, Ito, Hirayama, Yamamoto, Fujimoto, Kajita, Wakabayashi & Sobue, 2015). Following subthalamic

nucleus deep brain stimulation (STN-DBS), there is a discrepancy between the clinical/neurological assessment of speech, which mainly focuses on intelligibility, and the measurement of physiological markers connected to speech/voice dimensions. While a typical loss of intelligibility (of about 15 %) is found one year post-surgery, improvements in speech sub-components can also be observed, like for loudness (Tsuboi *et al.*, 2015; Tripoliti, Zrinzo, Martinez-Torres, Frost, Pinto, Foltynie, Holl, Petersen, Roughton, Hariz & Limousin, 2011). Lower preoperative speech intelligibility on medication, long disease duration and also medial left hemisphere active electrode contacts were reported as predictive factors of poor speech outcome following STN-DBS (Tripoliti, Limousin, Foltynie, Candelario, Aviles-Olmos, Hariz & Zrinzo, 2014). Stuttering and breathy voice have been attributed to disease progression and unrelated to STN-DBS (Tsuboi *et al.*, 2015), whereas strained voice and spastic-like dysarthria seem to be the result of current diffusion towards the cortico-bulbar tract (Pinto, Gentil, Krack, Sauleau, Fraix, Benabid & Pollak 2005; Tsuboi *et al.*, 2015; Atkinson-Clement, Maillet, LeBars, Lavenne, Redouté, Krainik, Pollak, Thobois & Pinto, 2017). Current diffusing towards cerebello-thalamo-cortical fibres was also considered to alter speech following STN-DBS (Tripoliti *et al.*, 2011; Tsuboi *et al.*, 2015; Atkinson-clement *et al.*, 2017). To improve speech outcome, neurostimulation parameters can be modulated, either by using low-frequency in conventional STN-DBS (Moreau *et al.*, 2011) or by using adaptive DBS (Little, Tripoliti, Beudel, Pogosyan, Cagnan, Herz, Bestmann, Aziz, Cheeran, Zrinzo, Hariz, Hyam, Limousin, Foltynie & Brown, 2016). Also, negative synergistic effects between STN-DBS and levodopa exist, and should be considered in treatment management (Pinto, Ferraye, Espesser, Fraix, Maillet, Guirchoum, Layani-Zemour, Ghio, Chabardès, Pollak & Debû, 2014). STN-DBS effects on speech in PD seem related to subjective cost over-valuation of the actual impairment: in a recent study, we demonstrated that vowel production effort was estimated easier to perform with STN-DBS alone, but harder when associated with levodopa (Atkinson-Clement, Cavazzini, Zénon, Legou, Witjas, Fluchère, Azulay, Baunez, Pinto & Eusebio, 2021). Finally, early STN-DBS may avoid side-effects on speech intelligibility (Pinto, Nebel, Rau, Espesser, R., Maillochon, Niebuhr, Krack, Witjas, Ghio, Cuartero, Timmermann, Schnitzler, Hesekamp, Meier, Müllner, Hälbig, Möller, Paschen, Paschen, Volkmann, Barbe, Fink, Becker, Reker, Kühn, Schneider, Fraix, Seigneuret, Kistner, Rascol, Brefel-Courbon, Ory-Magne, Hartmann, Wojtecki, Fradet, Maltête, Damier, Le Dily, Sixel-Döring, Benecke, Weiss, Wächter, Pinsker, Régis, Thobois, Polo, Houeto, Hartmann, Knudsen, Vidailhet, Schüpbach & Deuschl, 2023): in patients operated at an earlier stage of the disease, STN-DBS did not result in a consistent deterioration in blinded speech intelligibility assessment and patient-reported communication, as observed in studies of more advanced PD.

5.3 The importance of speech therapy

More than 90 % of the patients with PD complain about their low voice intensity and poor quality of communication, leading to a dramatic impairment in quality of

life and social isolation (Fox & Ramig, 1997). It has been shown that between 70 and 79 % of patients report a speech impairment (Hartelius & Svensson, 1994; Miller, Noble, Jones., Deane & Gibb, 2010). Among these patients, around 29 % stated that this symptom was the major one among all (Hartelius & Svensson, 1994). Despite that fact, only a few patients were referred to speech and language therapists, with percentages ranging from 1.6 to 4.4 % which seems to represent an underestimated need (Nesbitt & Thompson, 1995; Stewart, 2000; Trail, Fox, Ramig, Sapir, Howard & Lai, 2005; Hartelius & Svensson, 1994; Miller *et al.*, 2010). The question of patient referral to speech therapy remains a hot topic, even if such referral tends to increase in very recent years (Schalling, Johansson & Hartelius 2017) and depend largely on countries and areas. As part of multidisciplinary management (Goldman, Volpe, Ellis, Hirsch, Johnson, Wood, Aragon, Biundo, Di Rocco, Kasman, Iansek, Miyasaki, McConvey, Munneke, Pinto, St. Clair, Toledo, York, Todaro, Yarab & Wallock, 2024), speech therapy ought to be provided earlier and more frequently.

Despite insufficient evidence to support the efficacy of speech and language therapy for dysarthria in PD (Herd, Tomlinson, Deane, Brady, Smith, Sackley & Clarke, 2012), possibly also due to individual differences (Spurgeon, Clarke & Sackley, 2015), notable benefits have been observed (Atkinson-Clement, Sadat & Pinto, 2015) and intensive therapy demonstrated its ability to significantly and consistently improve dysarthria (Atkinson-Clement *et al.*, 2015; Ramig, Halpern, Spielman, Fox & Freeman, 2018).

6. *Future challenges*

Motor impairment in PD speech is not isolated. Dysarthria is also the result of auditory deficits, which appear to be quite specific in PD (Vitale, Marcelli, Allocca, Santangelo, Riccardi, Erro, Amboni, Pellecchia, Cozzolino, Longo, Picillo, Moccia, Agosti, Sorrentino, Cavaliere, Marciano & Barone, 2012; Vitale, Marcelli, Abate, Pianese, Allocca, Moccia, Spina, Barone, Santangelo & Cavaliere, 2016). This contributes probably to the degradation of auditory feedback processing (Arnold, Gehrig, Gispert, Seifried & Kell, 2014), as part of brain correlates underlying pathomechanisms of PD speech: 1) a striato-prefontal hypo-connectivity and dysfunctional self-monitoring mechanisms, underpinning the diminished motor drive of hypophonia; 2) a reduced external auditory feedback, affecting speech motor representations, and 3) a disturbed modulation of speech routines and affective prosody. These modifications could reflect either compensatory mechanisms or brain dysfunctions of PD speech (Arnold *et al.*, 2014). Higher-order language deficits are still underestimated in PD patients (e.g., Letanneux, Viallet, Velay & Pinto, 2021) and also require future investigation. In addition, rhythmic difficulties, emerging from probable generalised dysrhythmia ought to be considered to participate in the alteration of the temporal organisation of speech (Puyjarinet, Bégel, Gény, Driss, Cuartero, Kotz, Pinto & Dalla Bella, 2019; Puyjarinet, Bégel, Geny, Driss, Cuartero, Cochen De Cock, Pinto & Dalla Bella 2022).

From a theoretical point of view, the cross-linguistic approach to studying dysarthria is concerned with the question of whether disorders of dysarthric speech may differentially disturb the communication in a given language. Thus, dysarthric patients may fail to communicate important meaning differences in a certain language, but not in others in which these contrastive features are not distinctive. Studies examining the impact of speech difficulties from a cross-language perspective are increasingly numerous, providing arguments for language-universal or language-specific aspects of dysarthric speech. So far, cross-language studies suggest more similarities than differences in how dysarthria affects the speech of PD patients speaking different languages (e.g., Whitehill, Ma & Lee, 2003; Whitehill, 2010; Orozco-Arroyave, Hönig, Arias-Londoño, Vargas-Bonilla, Daqrouq, Skodda, Rusz & Nöth, 2016; Rusz, Hlavnička, Novotný, Tykalová, Pelletier, Montplaisir, Gagnon, Dušek, Galbiati, Marelli, Timm, Teigen, Janzen, Habibi, Stefani, Holzknecht, Seppi, Evangelista, Rassu, Dauvilliers, Högl, Oertel, St. Louis, Ferini-Strambi, Růžička, Postuma & Šonka, 2021; Verkhodanova, Coler, Jonkers, Timmermans, Maurits, De Jong & Lowie, 2022; Verkhodanova, Coler, Jonkers & Lowie, 2022). Despite these universals, García, De Leon, Tee, Blasi & Gorno-Tempini (2023) report that linguistic idiosyncrasies have also been found in PD speech (Kim & Choi, 2017; García *et al.*, 2023; García, Bocanegra, Herrera, Moreno, Carmona, Baena, Lopera, Pineda, Melloni, Legaz, Muñoz, Sedeño, Baez & Ibáñez, 2018; Møller, Høj, Østergaard, Wallentin & Højlund, 2023; García & Ibáñez, 2023).

In a recent work (Pinto, Cardoso, Atkinson-Clement, Guimarães, Sadat, Santos, Mercier, Carvalho, Cuartero, Oliveira, Welby, Frota, Cavazzini, Vigário, Letanneux, Cruz, Brulefert, Desmoulins, Pavão Martins, Rothe-Neves, Viallet & Ferreira, 2024), we aimed at identifying the relative contribution of acoustic variables to distinguish PD patients and controls who spoke varieties of two Romance languages, French (from France) and Portuguese (from Portugal). In a large cohort that included 129 PD patients and 124 healthy controls, we showed that acoustic variables that distinguished between speakers of the two languages were mainly related to phonation and voice quality. They contributed less than variables related to pneumo-phonic coordination and articulation rate at discriminating between PD patients and controls, confirming that respiration and diadochokinesis tasks seem the most appropriate to pinpoint signs of dysarthria, which are largely homogeneous and language-universal. In contrast, identifying language-specific variables was less conclusive with the speech tasks and acoustic variables studied here. Further research on that point is still required.

7. *Conclusive remarks*

Studying dysarthria starts often by identifying the communication deficits reported by the patients, and the consequences of speech impairment in their lives. It is of utmost importance to measure/document speech with a standardized assessment, early in the patient management, repeatedly across the disease progression, accompanied

by patient-reported assessments of voice/speech and/or communication deficits, and completed by systematic audiometric and higher-order cognitive-linguistic evaluations. Simply reporting global perceptual deficits or intelligibility decline fails to address important deficits that potentially contribute largely to speech impairment. Assessments that aim at providing a holistic evaluation are widely encouraged.

There is still a need for further research of dysarthria in PD, for a better understanding of its pathophysiology, treatment response, and relationship with other functions/symptoms, etc. Language-universal and language-specific dimensions also need to be further explored.

Acknowledgements

I wish to thank all my students and colleagues (neurologists, speech scientists, speech and language therapists, etc.) for their commitment to all our investigations and their outstanding contributions to our research field. I would like to thank also Dr Barbara Gili Fivela for the opportunity of submitting this point of view article on the fascinating topic we work on.

References

Ackermann, H., Ziegler, W. (1989). [Dysarthrophonia of Parkinson syndrome]. In *Fortschritte der Neurologie-Psychiatrie*, 57(4), 149-160. doi: 10.1055/s-2007-1000756

Ackermann, H. (2008). Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. In *Trends in Neurosciences,* 31, 265-272. doi: 10.1016/j.tins.2008.02.011

Arnold, C., Gehrig, J., Gispert, S., Seifried, C. & Kell, C.A. (2014). Pathomechanisms and compensatory efforts related to Parkinsonian speech. In *Neuroimage Clinical*, 4, 82-97. https://doi.org/10.1016/j.nicl.2013.10.016

Atkinson-Clement, C., Sadat, J. & Pinto, S. (2015). Behavioral treatments for speech in Parkinson's disease: meta-analyses and review of the literature. In *Neurodegenerative Disease Management*, 5(3), 233-248. https://doi.org/10.2217/nmt.15.16

Atkinson-Clement, C., Maillet, A., LeBars, D., Lavenne, F., Redouté, J., Krainik, A., Pollak, P., Thobois, S. & Pinto, S. (2017). Subthalamic nucleus stimulation effects on single and combined tasks in Parkinson's disease patients: a PET study. In *Brain Imaging and Behavior*, 11, 1139-1153. https://doi.org/10.1007/s11682-016-9588-4

Atkinson-Clement, C., Cavazzini, E., Zénon, A., Legou, T., Witjas, T., Fluchère, F., Azulay, J-P., Baunez, C., Pinto, S. & Eusebio, A. (2021). Subthalamic stimulation breaks the balance between distal and axial signs in Parkinson's disease. In *Scientific Reports*, 11, 21810. https://doi.org/10.1038/s41598-021-01386-0

Auzou, P. (2007). Définition et classifications des dysarthries. In auzou, p., rolland-maunoury, v., pinto, s., ozsancäk, c. (a cura di.), *Les Dysarthries*. Marseille: Solal, 308-323.

Braak, H., Del Tredici, K., Rüb, U., de Vos R.A.I., Jansen Steur, E.N.H. & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. In *Neurobiology of Aging,* 24(2), 197-211. https://doi.org/10.1016/S0197-4580(02)00065-9

Brabenec, L., Mekyska, J., Galaz, Z. & Rektorova, I. (2017). Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. In *Journal of Neural Transmission*, 124, 303-334. https://doi.org/10.1007/s00702-017-1676-0

Brain, W.R.B. (1965). *Speech disorders: aphasia, apraxia, and agnosia*. *2nd edition*. London: Butterworth.

Bunton, K., Kent, R.D., Duffy, J.R., Rosenbek, J.C. & Kent, J.F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. In *Journal of Speech, Language, and Hearing Research*, 50(6), 1481-1495. https://doi.org/10.1044/1092-4388(2007/102)

Canter, G.J. (1963). Speech Characteristics of Patients with Parkinson's Disease: I. Intensity, Pitch, and Duration. In *Journal of Speech and Hearing Disorders*, 28(3), 221-229. https://doi.org/10.1044/jshd.2803.221

Darley, F.L., Aronson, A.E. & Brown, J.R. (1969a). Differential diagnostic patterns of dysarthria. In *Journal of Speech and Hearing Research*, 12(2), 246-269. https://doi.org/10.1044/jshr.1202.246

Darley, F.L., Aronson, A.E. & Brown, J.R. (1969b). Clusters of deviant speech dimensions in the dysarthrias. In *Journal of Speech and Hearing Research*, 12(3), 462-496. https://doi.org/10.1044/jshr.1203.462

Darley, F.L., Aronson, A.E. & Brown, J.R. (1975). *Motor Speech Disorders*. Philadelphia: Saunders W.B.

Dashtipour, K., Tafreshi, A., Lee, J. & Crawley, B. (2018). Speech disorders in Parkinson's disease: pathophysiology, medical management and surgical approaches. In *Neurodegenerative Disease Management*, 8(5), 337-348. https://doi.org/10.2217/nmt-2018-0021

De Letter, M., Santens, P., De Bodt, M., Van Maele, G., Van Borsel, J. & Boon, P. (2007). The effect of levodopa on respiration and word intelligibility in people with advanced Parkinson's disease. In *Clinical Neurology and Neurosurgery*, 109(6), 495-500. https://doi.org/10.1016/j.clineuro.2007.04.003

Duffy, J.R. (2005). Defining, Understanding, and Categorizing Motor Speech Disorders. In Duffy, J.R. (a cura di), *Motor Speech Disorders – Substrates, Differential Diagnosis, and Management*. *2nd Edition*. Saint Louis: Elsevier Mosby, 3-16.

Duffy, J.R. (2007). Motor Speech Disorders: History, Current Practice, Future Trends and Goals. In Weismer, G. (a cura di), *Motor Speech Disorders*. San Diego: Plural Publishing, 7-56.

Duffy, J.R. (2013). Defining, Understanding, and Categorizing Motor Speech Disorders. In Duffy, J.R. (a cura di), *Motor Speech Disorders – Substrates, Differential Diagnosis, and Management. 3rd Edition*. Saint Louis: Elsevier Mosby, 3-14.

Fox, C.M., Ramig, L.O. (1997). Vocal sound pressure level and self-perception of speech and voice in men and women with idiopathic Parkinson disease. In *American Journal of Speech and Language Pathology*, 6(2), 85-94. https://doi.org/10.1044/1058-0360.0602.85

García, A.M., Bocanegra, Y., Herrera, E., Moreno, L., Carmona, J., Baena, A., Lopera, F., Pineda, D., Melloni, M., Legaz, A., Muñoz, E., Sedeño, L., Baez, S. & Ibáñez, A. (2018). Parkinson's disease compromises the appraisal of action mean-

ings evoked by naturalistic texts. In *Cortex*, 100, 111-126. https://doi.org/10.1016/j.cortex.2017.07.003

García, A.M., De Leon, J., Tee, B.L., Blasi, D.E. & Gorno-Tempini, M.L. (2023). Speech and language markers of neurodegeneration: A call for global equity. In *Brain*, 146(12), 4870-4879. https://doi.org/10.1093/brain/awad253

García, A.M., Ibáñez, A. (2023). On the replicability of action-verb deficits in Parkinson's disease. In *Cortex*, 158, 1-3. https://doi.org/10.1016/j.cortex.2022.10.008

Goldman, J.G., Volpe, D., Ellis, T.D., Hirsch, M.A., Johnson, J., Wood, J., Aragon, A., Biundo, R., Di Rocco, A., Kasman, G.S., Iansek, R., Miyasaki, J., McConvey, V.M., Munneke, M., Pinto, S., St. Clair, K.A., Toledo, S., York, M.K., Todaro, R., Yarab, N. & Wallock, K. (2024). Delivering Multidisciplinary Rehabilitation Care in Parkinson's Disease: An International Consensus Statement. In *Journal of Parkinson's disease*, 14(1), 135-166. https://doi.org/10.3233/JPD-230117

Golfinopoulos, E., Tourville, J.A. & Guenther, F.H. (2010). The integration of large-scale neural network modeling and functional brain imaging in speech motor control. In *NeuroImage*, 52(3), 862-874. https://doi.org/10.1016/j.neuroimage.2009.10.023

Grewel, F. (1957). Classification of dysarthrias. In *Acta Psychiatrica Scandinavica*, 32(3), 325-337. https://doi.org/10.1111/j.1600-0447.1957.tb09469.x

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics, 72, 43-53. https://doi.org/10.1007/BF00206237

Hartelius, L., Svensson, P. (1994). Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. In *Folia Phoniatrica et Logopaedica*, 46(1), 9-17. https://doi.org/10.1159/000266286

Head, H. (1926). *Aphasia and kindred disorders of speech*. New-York: Macmillan.

Herd, C.P., Tomlinson, C.L., Deane, K.H., Brady, M.C., Smith, C.H., Sackley, C.M. & Clarke, C.E. (2012). Comparison of speech and language therapy techniques for speech problems in Parkinson's disease. In *Cochrane Database of Systematic Reviews*, Issue 8, Art. No.: CD002814. https://doi.org/10.1002/14651858.CD002814.pub2

Hertrich, I., Dietrich, S. & Ackermann, H. (2020). The margins of the language network in the brain. *Frontiers in Communication,* 5, 519955. https://doi.org/10.3389/fcomm.2020.519955

Hickok, G., Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. In *Cognition*, 92(1-2), 67-99. https://doi.org/10.1016/j.cognition.2003.10.011

Kent, R.D. (2000). Research on speech motor control and its disorders: a review and prospective. *Journal of Communication Disorders,* 33(5), 391-427. https://doi.org/10.1016/S0021-9924(00)00023-X

Kent, R.D., Kent, J.F., Weismer, G. & Duffy, J.R. (2000). What dysarthrias can tell us about the neural control of speech. In *Journal of Phonetics*, 28(3), 273-302. https://doi.org/10.1006/jpho.2000.0122

Kim, Y., Choi, Y. (2017). A Cross-Language Study of Acoustic Predictors of Speech Intelligibility in Individuals With Parkinson's Disease. In *Journal of Speech, Language, and Hearing Research*, 60(9), 2506-2518. https://doi.org/10.1044/2017_JSLHR-S-16-0121

Kowalska-Taczanowska, R., Friedman, A. & Koziorowski, D. (2020) Parkinson's disease or atypical parkinsonism? The importance of acoustic voice analysis in differential diagnosis of speech disorders. In *Brain and Behavior*, 10(8), e01700. https://doi.org/10.1002/brb3.1700

Klostermann, F., Ehlen, F., Vesper, J., Nubel, K., Gross, M., Marzinzik, F., Curio, G. & Sappok, T. (2008). Effects of subthalamic deep brain stimulation on dysarthrophonia in Parkinson's disease. In *Journal of Neurology Neurosurgery and Psychiatry*, 79(5), 522-529. https://doi.org/10.1136/jnnp.2007.123323

Kröger, B.J., Kannampuzha, J., Eckers, C., Heim, S., Kaufmann, E. & Neuschaefer-Rube, C. (2012). The Neurophonetic Model of Speech Processing ACT: Structure, Knowledge Acquisition, and Function Modes. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (a cura di), *Cognitive Behavioural Systems. Lecture Notes in Computer Science*, vol 7403. Berlin, Heidelberg: Springer, 398-404. https://doi.org/10.1007/978-3-642-34584-5_35

Letanneux, A., Viallet, F., Velay, J-L. & Pinto, S. (2021) Altered inhibitory mechanisms in Parkinson's disease: evidence from Lexical Decision & Simple Reaction Time tasks. In *Frontiers in Human Neuroscience/Cognitive Neuroscience*, 15: 624026. https://doi.org/10.3389/fnhum.2021.624026

Liepmann, H.K. (1913). Motor aphasia, anarthria, and apraxia. In *Transactions of the 17th International Congress of Medicine*, Section XI, Part II, 97-106.

Little, S., Tripoliti, E., Beudel, M., Pogosyan, A., Cagnan, H., Herz, D., Bestmann, S., Aziz, T., Cheeran, B., Zrinzo, L., Hariz, M., Hyam, J., Limousin, P., Foltynie, T. & Brown, P. (2016). Adaptive deep brain stimulation for Parkinson's disease demonstrates reduced speech side effects compared to conventional stimulation in the acute setting. In *Journal of Neurology Neurosurgery and Psychiatry*, 87(12), 1388-1389. https://doi.org/10.1136/jnnp-2016-313518

Luchsinger, R., Arnold, G.E. (1965). *Voice-Speech-Language: Clinical communicology – Its physiology and pathology*. Belmont: Wadsworth Publishing Co.

Miller, N., Noble, E., Jones., D, Deane, K.H.O. & Gibb, C. (2010). Survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: patients' and carers' perspectives. In *International Journal of Language & Communication Disorders*, 46(2), 179-188. https://doi.org/10.3109/13682822.2010.484850

Møller, M.L.H., Høj, S.H., Østergaard, K., Wallentin, M. & Højlund, A. (2023). No selective action verb impairment in patients with Parkinson's disease: Evidence from Danish patients reading naturalistic texts, a Commentary on García *et al.*, 2018. In *Cortex*, 158, 176-180. https://doi.org/10.1016/j.cortex.2022.07.006

Moreau, C., Pennel-Ployart, O., Pinto, S., Plachez, A., Annic, A., Viallet, F., Destee, A. & Defebvre, L. (2011). Modulation of dysarthropneumophonia by low-frequency STN DBS in advanced Parkinson's disease. In *Movement Disorders*, 26(4), 659-663. https://doi.org/10.1002/mds.23538

Moreau, C., Pinto, S. (2019) Misconceptions about speech disorders in Parkinson's disease. In *Movement Disorders*, 34(10), 1471-1475. https://doi.org/10.1002/mds.27791

Mulroy, E., Stamelou, M. & Bhatia, K.P. (2019) How to approach a patient with parkinsonism – red flags for atypical parkinsonism. In *International Review of Neurobiology*, 149, 1-34. https://doi.org/10.1016/bs.irn.2019.10.001

Nesbitt, R., Thompson, R. (1995) Exploring interdisciplinary management of Parkinson's disease. In *International Journal of Language & Communication Disorders*, 30(S1), 414-414. https://doi.org/10.1111/j.1460-6984.1995.tb01723.x

Orozco-Arroyave, J.R., Hönig, F., Arias-Londoño, J.D., Vargas-Bonilla, J.F., Daqrouq, K., Skodda, S., Rusz, J. & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. In *The Journal of the Acoustical Society of America*, 139(1), 481-500. https://doi.org/10.1121/1.4939739

Parkinson, J. (1817) *An essay on the shaking palsy*. London: Sherwood, Neely and Jones.

Pinto, S., Gentil, M., Krack, P., Sauleau, P., Fraix, V., Benabid, A.L. & Pollak, P. (2005) Changes induced by levodopa and subthalamic nucleus stimulation on parkinsonian speech. In *Movement Disorders*, 20(11), 1507-1515. https://doi.org/10.1002/mds.20601

Pinto, S., Ghio, A. (2008). Les pathologies du contrôle moteur de la parole: contribution de l'étude des dysarthries et des dysphonies à la compréhension des mécanismes de la parole normale. In *Revue Française de Linguistique Appliquée*, XIII, 45-57. https://doi.org/10.3917/rfla.132.0045

Pinto, S, Ferraye, M, Espesser, R, Fraix, V., Maillet, A., Guirchoum, J., Layani-Zemour, D., Ghio, A., Chabardès, S., Pollak, P. & Debû, B. (2014). Stimulation of the pedunculopontine nucleus area in Parkinson's disease: effects on speech and intelligibility. In *Brain*, 137(10), 2759-2772. https://doi.org/10.1093/brain/awu209

Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R. & Sadat, J. (2017). A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. In *Journal of Phonetics*, 64, 156-167. https://doi.org/10.1016/j.wocn.2017.01.009

Pinto, S., Nebel, A., Rau, J., Espesser, R., Maillochon, P., Niebuhr, O., Krack, P., Witjas, T., Ghio, A., Cuartero, M-C., Timmermann, L., Schnitzler, A., Hesekamp, H., Meier, N., Müllner, J., Hälbig, T., Möller, B., Paschen, S., Paschen, L., Volkmann, J., Barbe, M., Fink, G., Becker, J., Reker, P., Kühn, A., Schneider, G.H., Fraix, V., Seigneuret, E., Kistner, A., Rascol, O., Brefel-Courbon, C., Ory-Magne, F., Hartmann, C.J., Wojtecki, L., Fradet, A., Maltête, D., Damier, P., Le Dily, S., Sixel-Döring, F., Benecke, P., Weiss, D., Wächter, T., Pinsker, M.O., Régis, J., Thobois, S., Polo, G., Houeto, J-L., Hartmann, A., Knudsen, K., Vidailhet, M., Schüpbach, M. & Deuschl, G. (2023) Results of a randomized clinical trial of speech after early neurostimulation in Parkinson's disease. In *Movement Disorders*, 38(2), 212-222. https://doi.org/10.1002/mds.29282

Pinto, S., Cardoso, R., Atkinson-Clement, C., Guimarães, I., Sadat, J., Santos, H., Mercier, C., Carvalho, J., Cuartero, M-C., Oliveira, P., Welby, P., Frota, S., Cavazzini, E., Vigário, M., Letanneux, A., Cruz, M., Brulefert, C., Desmoulins, M., Pavão Martins, I., Rothe-Neves, R., Viallet, F. & Ferreira, J.J. (2024). Do acoustic characteristics of dysarthria in people with Parkinson's disease differ across languages? In *Journal of Speech, Language and Hearing Research,* 67(9):2822-2841. doi: 10.1044/2024_JSLHR-23-00525

Poeppel, D., Emmorey, K., Hickok, G. & Pylkkänen., L. (2012). Towards a New Neurobiology of Language. In *Journal of Neuroscience,* 32(41), 14125–14131. https://doi.org/10.1523/JNEUROSCI.3244-12.2012

Price, C.J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading.In *NeuroImage,* 62(2), 816-847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Puyjarinet, F., Bégel, V., Gény, C., Driss, V., Cuartero, M-C., Kotz, S., Pinto, S. & Dalla Bella, S. (2019) Heightened orofacial, manual, and gait variability in Parkinson's disease results from a general rhythmic impairment. In *npj Parkinson's disease*, 5, 19, 1-7. https://doi.org/10.1038/s41531-019-0092-6

Puyjarinet, F., Bégel, V., Geny, C., Driss, V., Cuartero, M-C., Cochen De Cock, V., Pinto, S. & Dalla Bella, S. (2022) At-home training with a rhythmic video game lowers orofacial, manual, and gait variability in Parkinson's disease. In *Frontiers in Human Neuroscience/ Auditory Cognitive Neuroscience*, 16, 874032. https://doi.org/10.3389/fnins.2022.874032

Ramig, L., Halpern, A., Spielman, J., Fox, C. & Freeman, K. (2018). Speech treatment in Parkinson's disease: Randomized controlled trial (RCT). In *Movement Disorders*, 33(11), 1777-1791. https://doi.org/10.1002/mds.27460

Rusz, J., Tykalová, T., Klempíř, J., Čmejla, R. & Růžička, E. (2016). Effects of dopaminergic replacement therapy on motor speech disorders in Parkinson's disease: longitudinal follow-up study on previously untreated patients. In *Journal of Neural Transmission*, 123, 379-387. https://doi.org/10.1007/s00702-016-1515-8

Rusz, J., Tykalová, T., Ramig, L.O. & Tripoliti, E. (2021). Guidelines for Speech Recording and Acoustic Analyses in Dysarthrias of Movement Disorders. In *Movement Disorders*, 36(4), 803-814. https://doi.org/10.1002/mds.28465

Rusz, J., Tykalová, T., Salerno, G., Bancone, S., Scarpelli, J. & Pellecchia M.T. (2019). Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. In *Journal of Neurology*, 266(6), 1394-1404. https://doi.org/10.1007/s00415-019-09271-7

Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., Gagnon, J., Dušek, P., Galbiati, A., Marelli, S., Timm, P.C., Teigen, L.N., Janzen, A., Habibi, M., Stefani, A., Holzknecht, E., Seppi, K., Evangelista, E., Rassu, A.L., Dauvilliers, Y., Högl, B., Oertel, W., St. Louis, E.K., Ferini-Strambi, L., Růžička, E., Postuma, R.B. & Šonka, K. (2021). Speech Biomarkers in Rapid Eye Movement Sleep Behavior Disorder and Parkinson Disease. In *Annals of Neurology*, 90(1), 62-75. https://doi.org/10.1002/ana.26085

Schalling, E., Johansson, K. & Hartelius, L. (2017). Speech and Communication Changes Reported by People with Parkinson's Disease. In *Folia Phoniatrica et Logopaedica*, 69(3), 131-141. https://doi.org/10.1159/000479927

Selby, G. (1968). Parkinson's disease. In Vinken, P.J. & Bruyn, G.W. (Eds.), *Diseases of the basal ganglia*. Amsterdam: North Holland Publishing Company, 173-211.

Skodda, S., Grönheit, W., Mancinelli, N. & Schlegel, U. (2013). Progression of Voice and Speech Impairment in the Course of Parkinson's Disease: A Longitudinal Study. In Parkinson Disease, vol. 2013, Article ID 389195, 1-8. https://doi.org/10.1155/2013/389195

Skodda, S., Visser, W. & Schlegel, U. (2010). Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease. In *Journal of Neural Transmission*, 117, 197-205. https://doi.org/10.1007/s00702-009-0351-5

Spurgeon, L., Clarke, C.E., Sackley, C. (2015). Subjective Experiences of Speech and Language Therapy in Patients with Parkinson's Disease: A Pilot Study. In *Rehabilitation Research and Practice*, 839895. doi: 10.1155/2015/839895

Stewart, C. (2000). Speech-Language Therapy for Patients with Parkinson's Disease. In *Loss, Grief & Care*, 8(3-4), 151-155. https://doi.org/10.1300/J132v08n03_26

Tourville, J.A., Guenther, F.H. (2011). The DIVA model: A neural theory of speech acquisition and production. In *Language and Cognitive Processes*, 25, 952-981. https://doi.org/10.1080/01690960903498424

Trail, M., Fox, C., Ramig, L., Sapir, S., Howard, J. & Lai, E. (2005). Speech treatment for Parkinson's disease. In *NeuroRehabilitation*, 20(3), 205-221. doi: 10.3233/NRE-2005-20307

Tripoliti, E., Zrinzo, L., Martinez-Torres, I., Frost, E., Pinto, S., Foltynie, T., Holl, E., Petersen, E., Roughton, M., Hariz, M.I. & Limousin, P. (2011). Effects of subthalamic stimulation on speech of consecutive patients with Parkinson disease. In *Neurology*, 76(1), 80-86. https://doi.org/10.1212/WNL.0b013e318203e7d0

Tripoliti, E., Limousin, P., Foltynie, T., Candelario, J., Aviles-Olmos, I., Hariz, M.I. & Zrinzo, L. (2014). Predictive factors of speech intelligibility following subthalamic nucleus stimulation in consecutive patients with Parkinson's disease: Speech Intelligibility after STN-DBS. In *Movement Disorders*, 29(4), 532-538. https://doi.org/10.1002/mds.25816

Tsuboi, T., Watanabe, H., Tanaka, Y., Ohdake, R., Yoneyama, N., Hara, K., Nakamura, R., Watanabe, H., Senda, J., Atsuta, N., Ito, M., Hirayama, M., Yamamoto, M., Fujimoto, Y., Kajita, Y., Wakabayashi, T. & Sobue, G. (2015) Distinct phenotypes of speech and voice disorders in Parkinson's disease after subthalamic nucleus deep brain stimulation. In *Journal of Neurology Neurosurgery and Psychiatry*, 86(8), 856–64. https://doi.org/10.1136/jnnp-2014-308043

Verkhodanova, V., Coler, M., Jonkers, R. & Lowie, W. (2022). How expertise and language familiarity influence perception of speech of people with Parkinson's disease. In *Clinical Linguistics & Phonetics*, 36(2-3), 165-182. https://doi.org/10.1080/02699206.2021.2003433

Verkhodanova, V., Coler, M., Jonkers, R., Timmermans, S., Maurits, N., De Jong, B. & Lowie, W. (2022). A cross-linguistic perspective to classification of healthiness of speech in Parkinson's disease. In *Journal of Neurolinguistics*, 63, 101068. https://doi.org/10.1016/j.jneuroling.2022.101068

Vitale, C., Marcelli, V., Allocca, R., Santangelo, G., Riccardi, P., Erro, R., Amboni, M., Pellecchia, M.T., Cozzolino, A., Longo, K., Picillo, M., Moccia, M., Agosti, V., Sorrentino, G., Cavaliere, M., Marciano, E. & Barone, P. (2012). Hearing impairment in Parkinson's disease: Expanding the nonmotor phenotype. In *Movement Disorders*, 27(12), 1530-1535. https://doi.org/10.1002/mds.25149

Vitale, C., Marcelli, V., Abate, T., Pianese, A., Allocca, R., Moccia, M., Spina, E., Barone, P., Santangelo, G. & Cavaliere, M. (2016) Speech discrimination is impaired in parkinsonian patients: Expanding the audiologic findings of Parkinson's disease. In *Parkinsonism and Related Disorders*, 22(S1), S138-S143. https://doi.org/10.1016/j.parkreldis.2015.09.040

Wichmann, T. (2019). Changing views of the pathophysiology of Parkinsonism. In *Movement Disorders* 34(8), 1130-1143. https://doi.org/10.1002/mds.27741

Whitehill, T.L., Ma, J.K. & Lee, A.S. (2003). Perceptual characteristics of Cantonese hypokinetic dysarthria. In *Clinical Linguistics & Phonetics*, 17(4-5), 265-271. https://doi.org/10.1080/0269920031000080082

Whitehill, T.L. (2010). Studies of Chinese Speakers with Dysarthria: Informing Theoretical Models. In *Folia Phoniatrica et Logopaedica*, 62(3), 92-96. https://doi.org/10.1159/000287206

Ziegler, W., Ackermann, H., (2017). Subcortical contributions to motor speech: phylogenetic, developmental, clinical. In *Trends in Neurosciences* 40(8), 458-468. https://doi.org/10.1016/j.tins.2017.06.005

PARTE I

DISTURBI DEL PARLATO E DELLA FLUENZA

SPEECH AND FLUENCY DISORDERS

DOI: 10.17469/O2111AISV000004

DILETTA SCIANCALEPORE, VALENTINA LAVERMICOCCA,
TERESANNA ILLUZZI, SIMONE DIPACE, MARIA LUISA FIORELLA

Motor speech disorders and dysphagia in neurodegenerative diseases: clinical profiles compared

The aim of this study is to detect the alterations in the acoustic parameters of speech and swallowing difficulties in neurodegenerative diseases. 14 patients affected by Parkinson's Disease, 5 patients affected by Atypical Parkinsonism and 8 patients affected by Multiple Sclerosis were enrolled and underwent clinical and instrumental evaluation of speech and swallowing. The data obtained have been compared to those of a control group of 15 healthy people. The articulatory abilities have been assessed through maximum phonation time (MPT) and triangular voice space area (tVSA). Significant differences have been found between clinical and control group in MPT and tVSA. Regarding the swallowing assessment, parkinsonian syndromes showed worse results than Multiple Sclerosis in terms of bolus stagnation. In the patients group, speech and swallowing disorders are not similar in terms of neither severity nor onset. Noticeably, in our corpus dysphagia never represents the only disorder, as patients with dysphagia also show speech impairment.

Keywords: dysarthria, dysphagia, neurodegenerative disease, speech production.

1. *Introduction*

Dysarthria and dysphagia are common symptoms in neuromuscolar disease (Müller, Wenning, Verny, Mckee, Chaudhuri, Jellinger & Litvan., 2001; Nishio, Niimi, 2004; Perez-Lloret, Nègre-Pagès, Ojero-Senard, Damier, Destée, Tison, & Copark Study Group, 2012; Knuijt, Kalf, De Swart, Drost, Hendricks, Geurts., Van Engelen, 2013). 2013; Wang, Carter & Altman, 2020). Dysarthrias are speech disorders that result from neurologic impairments associated with weakness, slowness, or incoordination of the musculature used to produce speech. These speech disorders occur with considerable frequency in individuals with Parkinson disease, stroke, cerebellar disease, amyotrophic lateral sclerosis, multiple sclerosis, cerebral palsy, and traumatic brain injury (Duffy, 1995; McNeil, 1997; Kent, 2000). Neurogenic dysphagia is defined as an impairment of the swallowing ability that propels the food from oral cavity to the stomach and protects airway from food, liquids, and other substances in patients with neurological disease of various etiologies (Panebianco, Marchese-Ragona, Masiero & Restivo 2020). In literature, dysarthria appears to be a strong clinical indicator for the presence of dysphagia, and the severity of dysarthria seems to correspond to severity of dysphagia (Malandraki, Hind, & Gangnon, 2011; Ko, Chae & Cho, 2018; Wang *et al.*, 2020). The prevalence and severity of dysphagia varies depending on the type of dysarthria (Nishio *et al.*, 2004; Wang *et al.*, 2020).

The more severe conditions of dysphagia seem to be linked to dysarthria associated with bilateral impairment of upper or lower motor neurons such as flaccid, spastic, and mixed dysarthria. The idea that dysarthria and dysphagia could be correlated based on degree of severity is also supported in a 2010 study where the most severe dysarthria cases were associated with oropharyngeal dysphagia (Otapowicz, Sobaniec, Okurowska-Zawada, Artemowicz, Sendrowski, Kułak, Kuzia-Śmigielsk, 2010; Wang *et al.*, 2020). However, the correlation between dysarthria and oral-oropharyngeal dysphagia remains unclear (Wang, 2020). The aim of this study is to detect the presence of alterations in the acoustic parameters of the voice and swallowing function to outline clinical features of neurodegenerative diseases.

2. *Materials and methods*

The study included 14 people with Parkinson's disease (PD), 5 with atypical parkinsonism (PA) and 8 with multiple sclerosis (MS) enrolled at the at Ear, Nose and Throat Unit of University Hospital in Bari. Patients with other neurological or otolaryngological conditions that could affect swallowing or phono-articulation abilities were excluded from the study. Patients with Parkinson's disease were assessed using the Hoehn & Yahr clinical staging system (Hoen & Yahr, 1967), and patients with multiple sclerosis were assessed using the Expanded Disability Status Scale (Bushnik, 2011). The clinical group is 70 % composed by Parkinsonian patients and 30 % by Multiple Sclerosis patients. In the Parkinsonian group, the prevalence of the pathology in males is 58 % with a greater representation of the seventh decade of age. On average, the H&Y score is 2.8. Time since diagnosis (TSG) averages 7 years. The sample of people affected by Multiple Sclerosis is equally divided between the two sexes with a greater representation of the fourth decade of age. On average, the EDSS score is 4 (0.3). The TSG averages 20 years. Table 1 provides a summary of the demographic characteristics of the sample studied and the control group.

Table 1- *Demographic characteristics of clinical and control group*

	PD+PA	*MS*	*CG*
N (%)	19 (70)	8(30)	15
Mean age (±SD)	66.9 (±11.6)	50 (±11.2)	55 (±4.4)
Sex			
Male (%)	11 (58)	4 (50)	6 (40)
Female (%)	8 (42)	4 (50)	9 (60)
TSD (±SD)	7 (±3.6)	20 (±12.2)	
Stage of disease	2.8 (1.3) H&Y	4.1 (0.3) EDSS	

Each patient underwent clinical and instrumental swallowing evaluation and phonoarticulatory evaluation using acoustic analysis of voice. Control group subjects only underwent phonoarticulatory evaluation.

Statistical analysis of the data obtained was performed with the Software SPSS (IBM Corp, 2022): the T-student test, variance and covariance analysis were used.

2.1 Phono-articulatory evaluation

Voice recording was performed using a Samson Meteor Mic – USB Studio Condenser Microphone (frequency response of 20Hz-20kHz) placed 20 cm from the lips in a quiet room (30 dB background noise) and connected to the Audacity 2.1.2 software. The recording was done in mono channel with a sampling rate set at 44 kHz and 16 bits. The input volume was set to 70 %. The protocol proposed by Vizza, Tradigo and Mirarchi (2019) was used in the present study.

To evaluate the patients articulatory abilities, areas of the vowel triangles were analyzed. Each patient was instructed to perform a continuous and sustained vowel /a:/ to evaluate the Maximum Phonation Time. Patients able to pronounce a vowel for more than 5 seconds were asked to produce continuous and sustained vowels /a:/, /i:/, /u:/ for this duration. Subsequently, the triangular vowel space area of sustained vowels (tVSAv) was calculated using the F1 and F2 values of the sustained vowels /a:/, /i:/, /u:/. Subsequently subjects were asked to read a phonetically balanced text; the triangular vowel space area of semi-structured speech (tVSAs) was calculated using the average F1 and F2 values of three vowels /a:/, /i:/, /u:/ preceded and followed by plosive consonants in the text.

The vocal signal was analysed using Praat (Boersma, P. & Weenink, D., 2021). Areas were calculated using GeoGebra (Hohenwarter *et al.*, 2020).

2.2 Swallowing evaluation

Swallowing evaluation was carried out by clinical and instrumental assessment and was preceded by the administration of a self-report questionnaire. The Dysphagia Risk Assessment Scale (DRAS) questionnaire was used (Fukada, Kamakura & Kitaike, 2013; Cavallaro, Pontrelli & Sammali, 2022). The clinical assessment was carried out using the Bedside Swallowing Assessment Scale (BSAS), a scale that is useful in assessing the presence of the conditions necessary to start swallowing, the presence of defence mechanisms and the patient's ability to swallow using water testing. The instrumental assessment was then carried out by means of a fibroendoscopic examination of swallowing (FEES). The morphology and motility of the pharyngo-laryngeal structures and the search for possible salivary stagnation were assessed prior to the swallowing tests. The presence pre swallow spillage, penetration and aspiration, voluntary and reflex coughing, and post swallow stagnation in the valleculae and pyriform sinuses were observed at liquid (60 mL of water), semi-solid (yoghurt) and solid (dry biscuit) bolus test. Finally, the laryngeal adductory reflex (LAR) was assessed stimulating the arytenoid surface. The stagnation severity was assessed using the Yale Pharyngeal Residue Severity Rating Scale (Neubauer, Rademaker & Leder, 2015), which distinguishes between severe, moderate, mild, trace or absent stagnation. A score from 1 to 7 was obtained

on the Dysphagia Outcome and Severity Scale (DOSS) based on the impairment detected at the FEES.

3. *Results*

3.1 Swallowing evaluation results

Most patients included in the Parkinsonian group showed a mild to severe dysphagia with a DOSS score less than or equal to 5: 55 % of patients had a score equal to 5, 7 % of patients had a score equal to 4, 10.5 % of patients had a score equal to 3, 10.5 % of patients had a score equal to 2. The parkinsonian group showed an average DOSS score of 4.6, corresponding to a tendentially mild to moderate degree of dysphagia.In the Parkinsonian group, the average score obtained on the DRAS questionnaire is 14.

Clinical and instrumental evaluation of swallowing function showed pathological GAG-reflex in 53 % of cases, pathological voluntary cough in 32 % of cases, pathological LAR in 37 % of cases. Saliva stagnation was shown by 32 % of patients in the valleculae and by 21 % of patients in the pyriform sinuses. At liquid bolus tests, 42 % showed stagnation in valleculae and 44 % showed stagnation in pyriform sinuses. At semi-solid bolus tests, 56 % of patients showed stagnation in the valleculae and 50 % showed stagnation in the pyriform sinuses. At solid bolus tests 53 % of patients showed vallecular stagnation, 35 % showed stagnation in the piriform sinuses. No penetration was observed; aspiration was observed in 2 patients (11 %).

Patients included in the Multiple Sclerosis group showed a mild to moderate dysphagia: 29 % of patients had a score equal to 5, 5 % of patients had a score equal to 4. The average DOSS score was equal to 5.6, corresponding to mild degree of dysphagia. In the Multiple Sclerosis group, an average score of 12.2 was obtained from the DRAS self-assessment questionnaire.

Clinical and instrumental evaluation of swallowing function showed pathological GAG-reflex in 38 % of cases, pathological voluntary cough in 13 % of cases, pathological LAR in 38 % of cases. Saliva stagnation in the valleculae was observed in 13 % cases. At semi-solid bolus tests, 25 % of patients showed stagnation in the valleculae and 13 % showed stagnation in the pyriform sinuses. No stagnation was observed during liquid and bolus trials.

3.2 Speech evaluation results

The 100 % of patients included in the Parkinsonian group had a TOM score less than or equal to 4 showing a mild to severe dysarthria: 35 % of patients had a score equal to 4, 28 % of patients had a score equal to 3, 21 % of patients had a score equal to 2, 10 % of patients had a score equal to 1, 7 % of patients had a score equal to 0. The parkinsonian group showed an average TOM score of 2.7, corresponding to a tendentially moderate degree of dysarthria. The Maximum Phonation Time (MPT) has an average duration of 11s; 36 % of patients have a pathological phonatory

duration, ranging from 0s to 10s. The prevalent MPT is in the range of 9.1s -19s. The average value tVSAv is 145,739 Hz^2; the average tVSAs is 112,648 Hz^2.

In the Multiple Sclerosis group, 57 % of patients showed a mild to severe dysarthria: 75 % of them had a score equal to 4 and the remaining 25 % of patients had a score equal to 3. The average TOM scale score of 4 corresponding to a mild degree of dysarthria. MPT has an average duration of 10.83 s. The average value of tVSAv is 265,835 Hz^2; the average tVSAs is 157,094 Hz^2.

In the control group, MPT has an average duration of 23 s. The tVSAv average value is 260766 Hz^2 and the tVSAs average value is 289268 Hz^2.

Statistically significant difference was found between tVSAv and tVSAs in the clinical group ($p=0.04$) and borderline statistical significance between tVSAs in MS and control group ($p=0.05$). A borderline significant difference was found between the TMF of the MS patients and the control group ($p=0.05$). The analysis of covariance was used to study the trend of tVSAs, taking into account tVSAv as a covariate. A significant difference was found between the tVSAs values of controls compared to those of people with PD ($p=0.0002$), atypical PD ($p=0.0082$) and MS ($p=0.001$).

In the Parkinsonian group dysarthria and dysphagia coexisted in 83 % of cases, the remaining 17 % of patients only showed dysarthria. In the Multiple Sclerosis group 29 % of patients did not show either dysarthria or dysphagia, 38 % of patients only showed dysarthria and 14 % of patients only showed dysphagia. Swallowing and phono-articulatory disorders coexisted in 19 % of patients.

4. *Discussion*

The assessment of swallowing and phono-articulatory function made it possible to outline the clinical features of the pathologies under examination. Changes in the pneumophonic system, such as reduced respiratory support or glottic insufficiency, may explain the differences between the MPT in the clinical and in the control group.

Using tVSA for assessing speech intelligibility turned out to offer interesting results. The differences found between the control and clinical groups could be led to changes in the amplitude, strength and coordination of the movements required to produce phonemes. Within the clinical groups, it is observed a discrepancy between the values of the tVSA obtained from semi-structured speech and those obtained by extracting the formant values from sustained vowels. Indeed, the areas calculated on sustained vowels are significantly larger than those extracted from spontaneous speech. These results support the use of the vowel polygon extracted by more ecological tests to assess speech intelligibility. In the clinical group, most of the times both dysphagia and dysarthria were prevalent but speech and swallowing disorders were not similar either in terms of severity or onset. Atypical parkinsonism, which had the worst swallowing and phono-articulation scores, had the most severe clinical phenotype.

There was also a difference between the subjective ratings of the patients in the study and the ratings of the patients in the control group. Specifically, patients with Parkinson's disease and atypical parkinsonism showed a lower susceptibility to dysphagia, even when the dysphagia was overt. Multiple sclerosis patients, on the other hand, reported more frequently swallowing difficulties, even when undetectable by instrumental evaluation. As self-report is not always a reliable measure, it is recommended to assess swallowing abilities even in the absence of reported symptoms, especially in PD.

This finding corroborates the hypothesis that dysphagia should be always investigated in neurodegenerative patients affected by motor speech disorders.

5. *Conclusions*

It has been illustrated the swallowing and phonoarticulatory performance profiles in patients affected by Parkinson's Disease, Atypical Parkinsonism and Multiple Sclerosis, comparing the data obtained with those coming from a control group.

In the clinical group, most of the times both dysphagia and dysarthria were prevalent. The next goal is to continue the study by expanding the sample size, using more consistent data, which would allow to apply more valid statistics and produce results which are more representative of patients with neurodegenerative diseases.

Acknowledgments

The authors received the "Research Projects of National Relevance 2017 n. 2017JNKCYZ" funding for this work.

References

Audacity Team (2020). Audacity(R): Free Audio Editor and Recorder Computer application. Version 2.3.0. https://audacityteam.org/

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. Versione 6.1.40. https://www.praat.org/

Bushnik, T. (2011). Expanded Disability Status Scale. In: Kreutzer, J.S., DeLuca, J. & Caplan, B. (Eds.) Encyclopedia of Clinical Neuropsychology. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_1805

Cavallaro, G., Pontrelli, M. & Sammali, M. (2022). *Presbyphagia: the importance of an early diagnosis in the aging population.* In Journal of Gerontology and Geriatrics, 70, 237-243.

Duffy, J.R. (1995). *Motor speech disorders: Substrates, differential diagnosis, and management.* St. Louis: Mosby-Year Book.

Fukada, J., Kamakura, Y. & Kitaike, T. (2013). *Development of dysphagia risk assessment scale for elderly living at home.* In Yearbook of Journal of the Japan Society of Nursing Research.

Hoehn, M.M., Yahr, M.D. (1967). Parkinsonism: onset, progression and mortality. Neurology, 17(5), 427–442. https://doi.org/10.1212/wnl.17.5.427

Hohenwarter, M. (2020). Geogebra for Windows, Version 6.0. https://www.geogebra.org

IBM Corp. (2020). IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp. https://www.ibm.com/it-it/spss

Kent, R.D. (2000). *Research on speech motor control and its disorders: a review and prospective.* In Journal of Communication Disorders, 33(5), 391-427. https://doi.org/10.1016/s0021-9924(00)00023-x

Knuijt, S., Kalf, J.G., de Swart, B.J.M., Drost, G., Hendricks, H.T., Geurts, A.C.H. & van Engelen, B.G.M. (2013). *Dysarthria and dysphagia are highly prevalent among various types of neuromuscular diseases*. In Disability and Rehabilitation, 36(15), 1285-1289.

Ko, E.J., Chae, M. & Cho, R. (2018). *Relationship between swallowing function and maximum phonation time in patients with parkinsonism.* In Annals of Rehabilitation Medicine.

Müller, J., Wenning, G.K., Verny, M., McKee, A., Chaudhuri, K.R., Jellinger, K. & Litvan, I. (2001). *Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders*. In Archives of Neurology, 58(2), 259-264.

Malandraki, G., Hind, J. & Gangnon, R.L. (2011). *The Utility of Pitch Elevation in the Evaluation of Oropharyngeal Dysphagia: Preliminary Findings.* In American Journal of Speech-Language Pathology.

McNeil, M.R. (1997*). Clinical management of sensorimotor speech disorders.* New York: Thieme.

Neubauer, P.D., Rademaker, A.W. & Leder, S.B. (2015). *The Yale Pharyngeal Residue Severity Rating Scale: An Anatomically Defined and Image Based Tool.* In Dysphagia, 30, 521-528.

Nishio, M., Niimi, S. (2004). *Relationship between speech and swallowing disorders in patients with neuromuscular disease.* In Folia Phoniatrica et Logopaedica, 56(5), 291-304. https://doi.org/10.1159/000080066

Otapowicz, D., Sobaniec, W., Okurowska-Zawada, B., Artemowicz, B., Sendrowski, K., Kułak, W. & Kuzia-Śmigielska, J. (2010). *Dysphagia in children with infantile cerebral palsy*. In Advances in Medical Sciences, 55(2), 222-227.

Panebianco, M., Marchese-Ragona, R., Masiero, S. & Restivo, D.A. (2020). *Dysphagia in neurological diseases: a literature review.* In Neurological Sciences, 41(11), 3067-3073. https://doi.org/10.1007/s10072-020-04495-2

Perez-Lloret, S., Nègre-Pagès, L., Ojero-Senard, A., Damier, P., Destée, A., Tison, F. & COPARK Study Group (2012). *Oro-buccal symptoms (dysphagia, dysarthria, and sialorrhea) in patients with Parkinson's disease: preliminary analysis from the French COPARK cohort.* In European Journal of Neurology, 19(1), 28-37.

Vizza, P., Tradigo, G. & Mirarchi, D. (2019). *Methodologies of speech analysis for neurodegenerative diseases evaluation.* In International Journal of Medical Informatics, 122, 45-54.

Wang, B.J., Carter, F.L. & Altman, K.W. (2020). *Relationship between Dysarthria and Oral-Oropharyngeal Dysphagia: The present evidence.* In Ear, nose, & throat journal, Advance online publication. https://doi.org/10.1177/0145561320951647

DOI: 10.17469/O2111AISV000005

ALESSANDRA FERRERA, GLORIA GAGLIARDI

Acoustic biomarkers of Parkinson's Disease: a pilot study on the *Italian Parkinson's Voice and Speech* dataset

Beyond the most well-known motor symptoms, Parkinson's disease can also affect patients' speech and language abilities. This article presents an acoustic analysis of the *Italian Parkinson's Voice and Speech* dataset, comprising recordings of 65 speakers from southern Italy (28 patients, 22 elderly healthy controls, and 15 young healthy controls). The results reveal significant differences between pathological and control subjects across several metrics, including jitter, shimmer, HNR, and the Higuchi fractal dimension. Significant alterations were observed in vowel metrics, with a marked decrease in tVSA and VAI, and an increase in FCR. Consistent with existing literature, the study also confirms the documented reduction in pitch variability among patients. To complete the analysis, two additional variables (UPDRS scores and biological sex) were taken into account; notably, opposing trends between men and women emerged on duration measures (distribution of pauses, sound segments, and measures related to speech rate).

Keywords: Parkinson's disease, Parkinson, hypokinetic dysarthria, clinical linguistics, phonetics.

1. *Introduction*

Parkinson's Disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease (Rusz, Cmejla, 2011). From a neurological perspective, it is characterised by the progressive loss of dopaminergic cells in the substantia nigra. Alongside the primary motor signs of muscular rigidity, bradykinesia, rest tremor, and postural instability, the symptoms of the disease include various speech disorders, globally referred to as hypokinetic dysarthria, which can be grouped into three major components: dysphonia, articulatory problems, and dysprosody. Since several of these characteristics tend to manifest in the initial phases of the condition (Rusz, Cmejla, Ruzickova & Ruzicka, 2011), conducting research on them proves to be worthwhile, as it may eventually contribute to the development of diagnostic tools in the future.

Regarding the dysphonic traits, the voice of PD patients is often reported as being soft, breathy, harsh, hoarse, rough, and aspirate. These alterations in voice quality can be attributed to laryngeal dysfunctions that can affect patients, such as the condition of vocal fold bowing initially described by Hanson, Gerratt & Ward (1984).

As for the articulation of segmental features, a general trend towards articulatory undershooting has been observed, resulting from the effects of hypokinesia and rigidity on active articulators (Robertson, Hammerstad, 1996; Hunker, Abbs, 1990; Dworkin, Aronson, 1986). The reduction of articulatory gestures induces

alterations in acoustic vowel metrics: the triangular Vowel Space Area (tVSA) and the Vowel Articulation Index (VAI) tend to lower, while the Formant Centralisation Ratio (FCR) increases (Skodda, Visser & Schlegel, 2011a; Skodda, Grönheit & Schlegel, 2012). As regards to consonants, some studies have identified devoicing and spirantization patterns: incomplete closure of the articulators can result in plosives being produced as fricatives and can lead to the loss of the plosive element in affricates, while fricatives tend to lose the [+strident] feature (Antolík, Fougeron, 2013; Logemann, Fisher, 1981).

Concerning prosodic features, there is strong agreement in defining parkinsonian speech as monotonous in terms of pitch and intensity. There are greater discrepancies, however, regarding duration measures; while some studies have reported a significant decrease in verbal rate for PD patients (Peacher, 1950, cited in Cohen, 2003), others have not found significant differences compared to control subjects. Moreover, some other studies have even identified a tendency towards accelerated speaking rates towards the end of sentences; this phenomenon has been likened to the festination of gait and is commonly referred to as "oral festination" (Moreau, Ozsancak, Blatt, Derambure, Destee & Defebvre, 2007). As a final point, research on the rhythmic characteristics of PD has suggested a drift towards isosyllabic features, regardless of the rhythmic classification of the analysed language (Liss, White, Mattys, Lansford, Lotto, Spitzer & Caviness, 2009; Lowit, Marchetti, Corson & Kuschmann, 2018): this is seemingly a result of parkinsonian bradykinesia, leading the patient to lengthen stable articulatory configurations (vowels) at the expense of more dynamic ones (consonants).

Most of the above-mentioned features manifest cross-linguistically, as proved by the convergence of works on different languages (e.g., English, German, French, Mandarin, Czech, and Spanish). Some studies have also been conducted on Italian speakers, especially from Southern Italy. Various aspects of PD speech in patients from Salento (Puglia) have been explored by Gili Fivela and colleagues, including prosodic modulation (confirming the tendency towards monopitch in Gili Fivela, D'Apolito & Di Prizio, 2020a), higher vowels, and Italian geminates (Gili Fivela., Iraci, Grimaldi & Zmarich, 2015; Gili Fivela, Zmarich, 2005). An interesting outcome of their studies, derived from the combination of kinematic and acoustic analysis, is the hypothesis involving widest tongue movements as a compensatory mechanism for reduced labialisation in rounded vowels (Gili Fivela, D'Apolito & Di Prizio, 2020b).

Mirarchi, Vizza, Tradigo, Lombardo, Arabia & Veltri (2017) have measured VAI and VSA in PD patients from Cosenza (Calabria, Italy). As has been found in previous studies from other languages, these appear to be significantly reduced in PD patients compared to control subjects. Another research group, composed of Maffia, Pettorino, and their colleagues from the University of Naples "L'Orientale", successfully combined two rhythmic measures (%V and V-to-V) and found a significantly higher %V in the examined PD groups, even in drug-naive patients

(Pettorino, Busà & Pellegrino, 2016; Maffia, De Micco, Tessitore & De Meo, 2020; Maffia, De Micco, Pettorino, Siciliano, Tessitore & De Meo, 2021).

The present study hopes to follow this research tradition, with the aim of broadening the landscape of Italian works on hypokinetic dysarthria. To pursue this objective, we analysed the *Italian Parkinson's Voice and Speech* dataset, a fairly extensive corpus of recorded speech, fully available online (Dimauro, Di Nicola, Bevilacqua, Caivano & Girardi, 2017). The next section of this article (§ 2) provides a description of the composition of the dataset, with a focus on the specific tasks that have been selected for our purposes. In the following paragraph (§ 3), methods and results of our analysis will be presented. A short discussion of our findings will then follow in § 4.

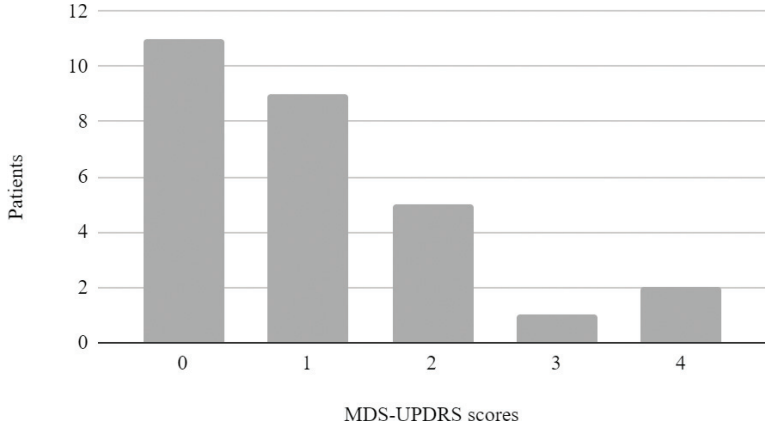
Although the previous literature has set certain expectations, which might have guided the interpretation of the results, we would describe our work as a mainly exploratory, not hypothesis-driven study. This choice is mainly related to the small size of our sample. Quantitative and statistical methods have been employed in all stages of the analysis, and the observation of statistical significance served the ultimate goal of uncovering potential new biomarkers for the disease.

2. Italian Parkinson's Voice and Speech*: description of the dataset*

All data were collected in 2017 in Bari (Puglia, Italy). A total of 65 speakers have been involved, divided into three groups: a first group of patients with Parkinson's disease (PPD) included 19 men and 9 women, whose ages ranged between 40-80 ($\mu = 67.2$); all of them came from the linguistic area of Bari, except from 1 male patient from Venice; a second group of 22 elderly healthy controls (EHC), 10 men and 12 women, age-ranging 60-77 ($\mu = 67.1$), all from the Bari area; a third group of young healthy controls (YHC), 13 men and 2 women, age range of 19-29 ($\mu = 20.8$), all of whom from Bari, except for 2 subjects from the Brindisi area. While we acknowledge the potential usefulness of including a third group (especially to understand age-related changes independent of pathology), our study specifically focused on the comparison between PPDs and EHCs.

All patients were receiving antiparkinsonian treatment and were in ON-state at the time of the experiment. The severity of the disease was examined by using the most common rating scales for motor examination: Hoehn & Yahr (HR) and MDS-UPDRS (Hoehn, Yahr, 1967; Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 2003). According to the HR scale, all patients received a score <4, except for two patients with stage 4, and one patient with stage 5. More details are provided by the authors (Dimauro *et al.*, 2017) about UPDRS scores, whose distribution is depicted by the bar chart in Fig. 1.

Figure 1 - *Distribution of PPD subjects based on UPDRS scores*



The corpus comprises recordings of various speech tasks, defined by the authors as the best protocol for automatic speech analysis. These include, listed in order:

a. Two readings of a phonemically balanced text, spaced by a pause (30 sec).
b. Execution of the syllable /pa/ (5 sec), pause (20 sec), execution of the syllable /ta/ (5 sec).
c. Two series of phonations for each vowel (/a/, /e/, /i/, /o/, /u/).
d. Reading of a list of phonemically balanced words.
e. Reading of a list of phonemically balanced sentences.

Among the available materials, recordings of tasks a., c., and e. appeared to be the most useful for our purposes and were selected for further analysis (see § 3).

As regards to task c., the participants were asked to inhale as much air as they could and to maintain the vowel as long as possible for the first phonation; in contrast, the second phonation was uniformly set to a duration of 5 seconds. Presented below are the text and sentences the participants read for points a. and e.

Phonemically balanced text

IL RAMARRO DELLA ZIA. Il papà (o il babbo come dice il piccolo Dado) era sul letto. Sotto di lui, accanto al lago, sedeva Gigi, detto Ciccio, cocco della mamma e della nonna. Vicino ad un sasso c'era una rosa rosso vivo e lo sciocco, vedendola, la volle per la zia. La zia Lulù cercava zanzare per il suo ramarro, ma dato che era giugno (o luglio non so bene) non ne trovava. Trovò invece una rana che saltando dalla strada finì nel lago con un grande spruzzo. Sai che fifa, la zia! Lo schizzo bagnò il suo completo rosa che divenne giallo come un taxi. Passava di lì un signore cosmopolita di nome Sardanapalo Nabucodonosor che si innamorò della zia e la portò con sé in Afghanistan.

List of phonemically balanced sentences

Oggi è una bella giornata per sciare. • Voglio una maglia di lana color ocra. • Il motociclista attraversò una strada stretta di montagna. • Patrizia ha pranzato a casa di Fabio. • Questo è il tuo cappello? • Dopo vieni a casa? • La televisione funziona? • Non posso aiutarti? • Marco non è partito. • Il medico non è impegnato.

As emphasized by the authors, the text is sufficiently long and is designed to test the speakers' resistance, as it contains a high concentration of complex and similar phonemes in close proximity. In the same way, the list of sentences was carefully chosen to engage all the muscles that are needed to produce the Italian language sounds.

3. *Methods and results*

The analysis was conducted in three main steps: the first involved the observation of acoustic vowel metrics, following the example of Skodda *et al.* (2012), Rusz, Cmejla, Tykalova, Ruzickova, Klempir, Majerova, Picmausova, Roth & Ruzicka (2013), and Sapir, Ramig, Spielman & Fox (2010). For the second part of the study, we relied on the use of the *DLBs Computational Pipeline*, a valuable tool for the automatic extraction of pathological speech features, originally developed by Gagliardi and Tamburini (2022). Thirdly, we reused the data calculated by Toye and Kompalli (2021) – especially regarding jitter, shimmer, and HNR measures– and applied a statistical analysis comparable to that of the previous stages. Further details about methods and results for each of these steps are provided in the following subsections (§ 3.1, § 3.2, and § 3.3).

3.1 Vowel description

Although the appearance of articulatory problems is often described as more tardive than dysphonia and dysprosody (Pinto, Ghio, Teston & Viallet, 2010; Rusz, Cmejla, Ruzickova & Ruzicka, 2011), alterations in acoustic vowel metrics have been proven to manifest early in the disease progression (Rusz *et al.*, 2013; Skodda *et al.*, 2012), and there is strong agreement on the direction of their shifts in PD patients.

To explore these metrics, the series of vowel executions (/a/, /e/, /i/, /o/, /u/; see task c. in § 2) were taken into account. All recordings have been manually analysed by collecting values of the first two formants (F1 and F2) in *Praat*; for every speaker, arithmetic mean values between the two executions of each vowel have then been calculated. Formant values were measured within a time window, according to criteria of stability and representativeness, rather than a fixed temporal length. The results obtained were statistically compared by applying the non-parametric test of Wilcoxon-Mann-Whitney in R.

Tab. 1 represents the mean values and standard deviations of F1 and F2 for both PPDs and EHCs, as well as the results of the statistical comparison (p-values) between the two groups. To highlight the most relevant results, only statistically significant outcomes are shown. Tab. 2, on the other hand, shows the results of the measurements of tVSA, VAI, and FCR[1], based on the mean formant frequencies of the corner vowels /a/, /i/, /u/.

[1] For further details and formulas of tVSA, VAI, and FCR, see Skodda *et al.* (2012) and Sapir *et al.* (2010).

Table 1 - *Means and st. deviations of formant frequencies in PD speakers and elderly controls (SD values for each measure in brackets). P-values in the right column mark the degree of statistical significance according to the following rules: * if p-value ≤ 0.05; ** if p-value ≤ 0.01; *** if p-value ≤ 0.001; **** if p-value ≤ 0.0001*

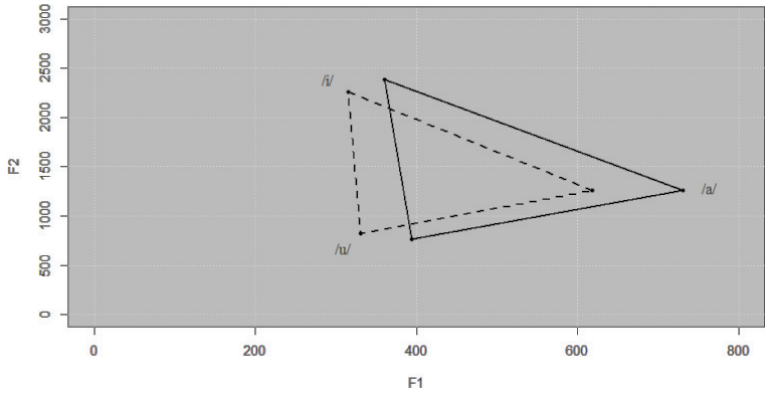
	PPD		*EHC*		*p-value*
/a/ F1 Mean	618.3746	(± 219.3455)	730.7576	(± 155.1223)	0.04236 *
/e/ F2 Mean	1829.374	(± 257.5197)	2033.51	(± 191.1282)	0.003334 **
/i/ F1 Mean	315.274	(± 79.84552)	359.9481	(± 59.97617)	0.002358 **
/u/ F1 Mean	331.0825	(± 62.36161)	394.0553	(± 49.10942)	0.000607 ***

Table 2 - *Triangular Vowel Space Area, Vowel Articulation Index, and Formant Centralisation Ratio for PPD and EHC*

	tVSA	*VAI*	*FCR*
PPD	209359.2	1.053374	0.94933
EHC	280145.1	1.119967	0.892883

As expected, the Vocalic Articulation Index (VAI) and tVSA are lower in PPDs compared to non-pathological subjects, confirming a tendency towards vowel centralisation in pathological subjects; conversely, the Formant Centralisation Ratio is higher in patients than in controls. A graphical comparison between triangular Vowel Space Areas for both the PPD and the EHC groups is provided in Fig. 2: as shown in the graph, the patients' area is visibly reduced compared to that of the EHC group, as well as the global mean value of F1.

Figure 2 - *Mean triangular Vowel Space Areas (tVSA) for the PPD group (dotted lines) and the EHC group (continuous lines)*



In order to refine the analysis, the additional variable of the speakers' sex has been taken into account. Tab. 3 shows the results of these further measurements, divided between male and female speakers. As in Tab. 1, p-values result from the statistical test of Wilcoxon-Mann-Whitney, comparing mean values of male and female patients to

the respective subgroups of elderly healthy controls of the same sex. Once again, only statistically significant results are presented.

Table 3 - *Means and standard deviations of formant values, divided between men and women. P-values result from the statistical comparison of PD subgroups with their respective controls of the same sex*

	Men			*Women*		
	PPD	*EHC*	*p-value*	*PPD*	*EHC*	*p-value*
/a/ F1 Mean	508.5833 (±161.222)	711.1058 (±121.3202)	0.002009 **	850.1561 (±121.778)	747.1341 (±182.3518)	0.1694
/a/ F2 Mean	1183.466 (±147.123)	1219.147 (±79.5595)	0.3077	1420.925 (±137.7207)	1289.934 (±122.8868)	0.04907 *
/e/ F1 SD	23.5133 (±18.22045)	13.5341 (±11.8644)	0.09439	7.1427 (±5.4647)	34.0311 (±27.7193)	0.04089 *
/e/ F2 Mean	1726.63 (±124.8017)	1892.996 (±151.1217)	0.00568 **	2046.28 (±332.8249)	2150.605 (±134.7093)	0.8078
/i/ F1 Mean	276.0183 (±33.9081)	350.2001 (±67.1847)	0.0000657 ****	398.1473 (±86.8523)	368.0714 (±54.9163)	0.6016
/i/ F2 Mean	2175.29 (±190.6583)	2300.147 (±149.8447)	0.04455 *	2444.188 (±407.8353)	2446.333 (±405.4428)	0.5538
/u/ F1 Mean	308.1545 (±54.9667)	383.7472 (±56.5358)	0.001665 **	379.486 (±49.4267)	402.6453 (±42.5673)	0.3451

The table reveals some sex-related differences: among men, in addition to the differences highlighted in Table 1, there is a further alteration in the second formant of /e/. For women, the only significant alterations pertain to the second formant of /a/ and the standard deviation of the first formant of /i/.

3.2 DLBs Computational Pipeline

As previously stated (§ 3), the second part of the analysis relied on the use of a pipeline for the automatic extraction of Digital Linguistic Biomarkers (DLBs), namely a set of acoustic and linguistic metrics that could potentially aid the identification of pathological speech (Gagliardi, 2023). The *Pipeline* has proved to be a promising tool for the detection of other pathologies (e.g., Mild Cognitive Impairment; see Gagliardi, Tamburini, 2021) and currently calculates a wide range of features, encompassing acoustic, rhythmic, lexical, and syntactic measures, as well as readability and linguistic inquiry and word count features. Coherently with our objectives and with the nature of the available data, solely the acoustic features were considered in this study. These are listed in Calzà, Gagliardi, Rossini Favretti &Tamburini (2021) and can be briefly described as follows:

– SILMEAN, SILMEDIAN, SILSD: mean, median and standard deviation of silent segments duration.
– SPEMEAN, SPEMEDIAN e SPESD: mean, median and standard deviation of speech segments duration (excluding silent segments).

- TRVSD: temporal regularity of voiced segment durations.
- VR: verbal rate, including silent segments.
- TPR: transformed phonation rate, calculated as arcsin√PR, where PR (phonation rate) is the ratio between total phonation time (speech time without pauses) and total locution time (speech time including pauses).
- SPT (Standardised phonation time): standardised phonation time, calculated as the ratio between the number of words and the duration of phonation (excluding silent segments).
- SPR (Standardised pause rate): ratio between words and silences in the audio sample.
- RMSEM e RMSESD: mean and standard deviation of the root mean square energy.
- PITCHM e PITCHSD: mean and standard deviation of pitch.
- SPCENTRM e SPCENTRSD: mean and standard deviation of spectral centroid, an acoustic correlate of the perceptual brightness of sounds.
- HFractDM e HFractDSD: mean and standard deviation of the Higuchi fractal dimension.

For this part of the analysis, tasks a. (two readings of a phonemically balanced text) and e. (reading of a list of phonemically balanced sentences) were selected. All data extracted from the corpus were adapted to fulfil the pipeline's requirements: audio files were converted to .wav/44.1KHz/16bit/Mono, and orthographic transcriptions (.txt, Unicode UTF-8) were provided.

Tables below present aggregate data for the PPD and EHC groups[2]. Measurements were carried out separately for each task; as for the other tables further down in this article, the degree of significance is indicated according to the same rules as Tab. 1. Notably, some of the metrics show a significant difference across all the tasks: SPEMEAN, SPEMEDIAN, SPT, PITCHSD, SPCENTRSD, FractDMEAN e HFractDSD (Tab. 4-6).

Table 4 - *Means, st. deviations and statistical comparison (PPDs vs EHCs) for the 1st text reading task*

	PPD		*EHC*		*p-value*
SPEMEAN	1.425933	(±0.55628)	0.982076	(±0.736492)	0.0362 *
SPESD	0.917477	(±0.33362)	0.642534	(±0.407299)	0.0223 *
SPEMEDIAN	1.263111	(±0.582036)	0.861851	(±0.72917)	0.0314 *
VR	2.154168	(±0.704509)	2.525489	(±0.437549)	0.0449 *
SPT	3.794013	(±3.247913)	8.64725	(±5.978701)	0 ****
PITCHM	148.4479	(±28.42092)	166.1551	(±28.08773)	0.0442 *
PITCHSD	23.95979	(±5.579796)	37.33001	(±8.963386)	0 ****
SPCENTRM	1916.485	(±823.4478)	2044.991	(±515.6468)	0.0391 *
SPCENTRSD	1560.532	(±428.6954)	1286.021	(±282.7223)	0.0346 *
HFractDMEAN	1.624539	(±0.090066)	1.829976	(±0.049112)	0 ****
HFractDSD	0.154929	(±0.024546)	0.085443	(±0.015314)	0 ****

[2] Only statistically significant results are shown.

Table 5 - *Means, st. deviations and statistical comparison (PPDs vs EHCs) for the 2nd text reading task*

	PPD		*EHC*		*p-value*
SPEMEAN	1.833724	(±0.530703)	1.221547	(±0.869531)	0.0112 *
SPESD	1.130024	(±0.30339)	0.812094	(±0.499486)	0.0117 *
SPEMEDIAN	1.699225	(±0.696153)	1.088312	(±0.902443)	0.0099 **
SPT	3.524875	(±0.573903)	7.606606	(±5.140292)	0.0012 **
PITCHM	145.1231	(±28.43825)	165.7067	(±27.40162)	0.0196 *
PITCHSD	22.62279	(±5.935683)	37.76195	(±8.418174)	0 ****
SPCENTRSD	1568.869	(±409.4062)	1276.227	(±279.6976)	0.0277 *
HFractDMEAN	1.623039	(±0.086177)	1.824355	(±0.048178)	0 ****
HFractDSD	0.155023	(±0.025143)	0.085367	(±0.016201)	0 ****

Table 6 - *Means, st. deviations and statistical comparison (PPDs vs EHCs) for the sentence reading task*

	PPD		*EHC*		*p-value*
SPEMEAN	1.709863	(±0.446501)	1.209864	(±0.812239)	0.01836 *
SPEMEDIAN	1.739127	(±0.55902)	1.170601	(±0.897708)	0.02078 *
VR	1.980705	(±0.492023)	2.37204	(±0.40088)	0.004044 **
SPT	3.044265	(±0.6706)	5.987555	(±3.285112)	6.414e-05 ****
PITCHSD	23.01193	(±6.222952)	32.20312	(±8.890534)	0.0002056 ***
SPCENTRSD	1697.961	(±585.1265)	1214.359	(±286.0436)	2.264e-05 ****
HFractDMEAN	1.641649	(±0.094701)	1.829654	(±0.051397)	3.491e-10 ****
HFractDSD	0.153559	(±0.023903)	0.077639	(±0.017954)	1.296e-11 ****

As in § 3.2, it was deemed useful to investigate some additional variables to assess their potential impact on the results. More specifically, we chose to consider the biological sex of the participants and, regarding patients, their score on a UPDRS rating scale. In Tab. 7-9, the subgroups of male and female PPDs are compared to respective controls of the same sex. These tables show opposing trends for male and female patients across some of the metrics: for example, SILMEAN, SILSD, and SILMEDIAN values prove to be higher in male patients and lower in female patients (compared to respective sex-matched controls), whereas SPEMEAN, SPEMEDIAN, and TPR values decrease in men and increase in women. Conversely, the Parkinson's effect is consistent across genders for other measures, such as PITCHSD and HFractDM (lower in both PPD groups), or SPCENTRSD and HFractDSD (always higher in PPDs than in controls).

Table 7 - *Male and female PPDs compared to their respective controls on the first text reading task*

	Men			*Women*		
	PPD	*EHC*	*p-value*	*PPD*	*EHC*	*p-value*
SIL MEAN	1.1734 (±0.4947)	0.9665 (±0.4888)	0.0071 **	1.0339 (±0.312)	1.8299 (±0.8279)	0.0029 **
SIL SD	0.5979 (±0.5156)	0.3668 (±0.3825)	0.0178 *	0.5058 (±0.3641)	1.1557 (±0.6358)	0.0084 **
SIL MEDIAN	1.0246 (±0.3466)	0.9291 (±0.3716)	0.0654	0.9111 (±0.167)	1.5475 (±0.7991)	0.0114 *
SPE MEAN	1.3003 (±0.5296)	1.6387 (±0.6516)	0.0293 *	1.72437 (±0.5324)	0.5282 (±0.31431)	0.0008 ***
SPE SD	0.8749 (±0.342)	0.9903 (±0.3326)	0.1164	1.0186 (±0.3099)	0.4018 (±0.2459)	0.001 ***
SPE MEDIAN	1.1235 (±0.5368)	1.5239 (±0.6773)	0.0249 *	1.5946 (±0.5822)	0.407 (±0.2218)	0.0005 ***
VR	2.2041 (±0.6078)	2.64004 (±0.2913)	0.0357 *	2.0355 (±0.9332)	2.4416 (±0.5032)	0.2197
TPR	0.8998 (±0.1896)	1.0083 (±0.2567)	0.0089 **	1.0349 (±0.2054)	0.5563 (±0.2159)	0.0009 ***
SPT	4.263 (±3.7757)	4.5526 (±2.6819)	0.383	2.6802 (±0.7133)	11.5067 (±6.1643)	0.0001 ****
PITCH SD	23.0117 (±5.7587)	39.3436 (±9.6425)	0.0002 ***	26.2115 (±4.7045)	36.1734 (±7.8735)	0.0102 *
SPCENTRSD	1568.122 (±362.359)	1123.302 (±200.7114)	0.0065 **	1542.506 (±586.8935)	1388.472 (±291.8508)	0.3839
HFractD MEAN	1.586 (±0.0742)	1.8023 (±0.0334)	0 ****	1.7161 (±0.0488)	1.8477 (±0.0519)	0.0003 ***
HFractD SD	0.1624 (±0.0203)	0.089 (±0.0129)	0 ****	0.1372 (±0.0258)	0.0837 (±0.018)	0.0004 ***

Table 8 - *Male and female PPDS compared to their respective controls on the 2nd text reading task*

	Men			*Women*		
	PPD	*EHC*	*p-value*	*PPD*	*EHC*	*p-value*
SIL MEAN	0.9683 (±0.1554)	0.8279 (±0.2419)	0.0074 **	0.9342 (±0.3269)	1.6359 (±0.6434)	0.009 **
SIL SD	0.34999 (±0.1146)	0.2915 (±0.2581)	0.0247 *	0.5508 (±0.549156)	1.0049 (±0.5903)	0.0641
SIL MEDIAN	0.9366 (±0.1594)	0.7947 (±0.1873)	0.0099 **	0.8449 (±0.1758)	1.3812 (±0.5329)	0.0192 *
SPE MEAN	1.7186 (±0.4746)	1.9449 (±0.7665)	0.1203	2.16002 (±0.588)	0.67902 (±0.4452)	0.0012 **
SPE SD	1.0708 (±0.2767)	1.1856 (±0.4403)	0.2285	1.2979 (±0.338)	0.53215 (±0.3363)	0.0013 **

	Men			*Women*		
	PPD	*EHC*	*p-value*	*PPD*	*EHC*	*p-value*
SPE MEDIAN	1.5739 (±0.69899)	1.8224 (±0.8903)	0.1631	2.0543 (±0.6035)	0.5377 (±0.3642)	0.0012 **
TPR	1.0127 (±0.0802)	1.0812 (±1.9425)	0.0152 **	1.1186 (±0.1816)	0.6397 (±0.2594)	0.0031 **
SPT	3.6803 (±0.489245)	3.9885 (±1.4963)	0.2685	3.0844 (±0.6068)	10.32015 (±5.2602)	0.0005 ***
SPR	6.1178 (±1.8423)	6.7308 (±1.9425)	0.2957	6.5906 (±2.4799)	5.1473 (±1.1319)	0.0482 *
PITCH SD	20.9433 (±5.2527)	37.8978 (±8.8225)	0.0003 ***	27.3814 (±5.4744)	37.66004 (±8.4977)	0.015 *
SPCENTRSD	1575.213 (±327.533)	1084.074 (±191.7497)	0.0025 **	1550.895 (±627.4434)	1420.341 (±250.262)	0.4108
HFractD MEAN	1.5838 (±0.0589)	1.8002 (±0.0286)	0 ****	1.7341 (±0.0396)	1.8425 (±0.0528)	0.0024 **
HFractD SD	0.1635 (±0.0191)	0.0857 (±0.0097)	0 ****	0.1309 (±0.0258)	0.0851 (±0.0202)	0.0024 **

Table 9 - *Male and female PPDs compared to their respective controls on the sentence reading task*

	Men			*Women*		
	PPD	*EHC*	*p-value*	*PPD*	*EHC*	*p-value*
SIL SD	0.4204 (±0.3124)	0.4094 (±0.3386)	0.7512	0.3604 (±0.339)	0.8938 (±0.6182)	0.02013 *
SPE MEAN	1.6630 (±0.3796)	1.7845 (±0.8259)	0.491	1.8094 (±0.5809)	1.5343 (±0.6182)	0.001524 **
SPE SD	0.8197 (±0.2056)	0.8621 (±0.2433)	0.5604	0.8679 (±0.1866)	0.7789 (±0.2691)	0.03142 *
SPE MEDIAN	1.7032 (±0.5202)	1.8197 (±0.9325)	0.458	1.8154 (±0.6657)	0.5881 (±0.4791)	0.001064 **
VR	2.0749 (±0.3662)	2.5209 (±0.3056)	0.002857 **	1.7805 (±0.6754)	2.2604 (±0.4388)	0.1153
TPR	0.9211 (±0.0914)	0.9704 (±0.25424)	0.2198	0.98797 (±0.1199)	0.6787 (±0.2255)	0.004096 **
SPT	3.3033 (±0.5277)	4.3815 (±2.0571)	0.2409	2.4939 (±0.6294)	7.19207 (±3.5848)	0.000111 ***
PITCH SD	21.0772 (±5.7944)	36.5242 (±9.6271)	0.000309 ***	27.1233 (±5.2554)	28.9623 (±7.04081)	0.5208
SPCENTRSD	1612.838 (±363.1613)	1055.049 (±141.0105)	0.00041 ***	1878.847 (±413.509)	1333.841 (±312.994)	0.007303 **
HFractD MEAN	1.6012 (±363.1613)	1.807 (±0.0349)	6.401e-07 ****	1.7276 (±0.0612)	1.8466 (±0.0564)	0.001524 **
HFractD SD	0.1599 (±0.0166)	0.0768 (±0.0127)	6.401e-07 ****	0.1402 (±0.0320)	0.0783 (±0.0216)	0.000714 ***

To evaluate the variation according to UDRPS scores (Tab. 10), PPDs were divided into three subgroups: patients with mild symptoms (scoring 0 or 1), moderate (2-3), and severe (UPDRS>3) symptoms; it should be noted that the comparison between the three groups was only possible for the first reading of the text, as patients with severe symptoms were unable to perform the second text reading and sentence reading task. The results from Tab. 10 show that, for many of the metrics, the difference becomes significant precisely in the transition to the most advanced stage of the disease.

Table 10 - *Comparison between UPDRS subgroups of patients, based on the 1st text reading task*

	Mild PD	*Moderate PD*	*Severe PD*	*p-values*		
				Mild vs Mod.	*Mod. vs Sev.*	*Mild vs Sev.*
SIL MEAN	1.1143 (±0.4186)	0.98898 (±0.3337)	1.9079 (±0.5566)	0.1276	0.0105 *	0.0367 *
SIL SD	0.5265 (±0.3832)	0.4406 (±0.4020)	1.4590 (±0.8361)	0.0945	0.0087 **	0.0393 *
SIL MEDIAN	0.9928 (±0.3237)	0.9057 (±0.1601)	1.4402 (±0.1912)	0.3605	0.0185 *	0.0216 *
SPE MEAN	1.4651 (±0.5571)	1.5828 (±0.5342)	0.6890 (±0.0093)	0.3605	0.0185 *	0.0216 *
SPE SD	0.9713 (±0.358)	0.9104 (±0.2563)	0.4723 (±0.0004)	0.4091	0.032 *	0.0168 *
SPE MEDIAN	1.2816 (±0.5696)	1.4748 (±0.5975)	0.5490 (±0.0166)	0.2212	0.0123 *	0.0256 *
TRV SD	0.4372 (±0.183)	0.5436 (±0.4058)	0.6865 (±0.0426)	0.3836	0.0593	0.032 *
VR	2.3026 (±0.544)	2.0959 (±0.8586)	0.8989 (±0.2011)	0.4091	0.032 *	0.0168 *
TPR	0.9376 (±0.1804)	1.0272 (±0.2045)	0.6165 (±0.0708)	0.2576	0.0139 *	0.0244 *
SPT	4.3412 (±4.0073)	2.8209 (±0.7136)	2.6759 (±0.0674)	0.0391 *	0.359	0.0742
SPR	5.1310 (±1.8984)	4.4894 (±2.0831)	1.8244 (±0.0203)	0.2688	0.0304 *	0.0086 *
PITCH SD	22.4547 (±5.1725)	26.3035 (±5.0885)	28.6058 (±10.2910)	0.0385 *	0.4629	0.1225
HFractD MEAN	1.6183 (±0.0884)	1.6745 (±0.0838)	1.5317 (±0.0713)	0.0834	0.0202 *	0.0856

3.3 Jitter, Shimmer, HNR

Since dysphonia has been shown to be among the initial indicators of speech impairment in PD (see Pinto *et al.*, 2010), exploring measures related to voice quality represents a particularly interesting research avenue. In 2021, Toye and Kompalli extracted some of these features (including jitter variants, shimmer variants, and

HNR[3]) from vowel execution recordings of the *Italian Parkinson's Voice and Speech* corpus; since the results of their measurement were made publicly available[4], it was possible to reuse them to supplement the present study. A statistical analysis was thus conducted on these data, following the same approach as the previous subsections: means and standard deviations were calculated for PPDs and EHCs, and the two groups were later compared by using the statistical non-parametric test of Wilcoxon-Mann-Whitney. Tables below illustrate the results obtained for each vowel on jitter measures (Tab. 11), shimmer measures (Tab. 12), and Harmonic-to-Noise Ratio (Tab. 13). Overall, the differences between the two groups were highly significant for all vocalizations in all shimmer and HNR measures, while, on the other hand, jitter measures showed statistical significance only for the three corner vowels /a/, /i/, and /u/.

Table 11 - *Jitter measures on vowel executions: means, standard deviations and statistical comparison (p-values) between PPDs and EHCs*

		Local Jitter	*Local Absolute Jitter*	*Rap Jitter*	*Ddp Jitter*
/a/	*Mean PPD*	0.005624 (±0.008486)	3.81e-05 (±0.0000643)	2.93e-03 (±0.005012)	0.008776 (±0.015036)
	Mean EHC	0.005838 (±0.003632)	3.94e-05 (±0.0000243)	0.00309 (±0.002142)	0.00927 (±0.006426)
	p-value	0.0328 *	0.0277 *	0.02155 *	0.02155 *
/e/	*Mean PPD*	0.006436 (±0.011325)	4.41e-05 (±0.0000844)	3.33e-03 (±0.006461)	0.009977 (±0.019384)
	Mean EHC	0.005312 (±0.004151)	3.61e-05 (±0.0000251)	0.002794 (±0.002399)	0.008382 (±0.007197)
	p-value	0.4352	6.01e-01	0.2894	0.2894
/i/	*Mean PPD*	0.003832 (±0.001592)	2.44e-05 (±0.0000116)	1.89e-03 (±0.000879)	0.005679 (±0.002636)
	Mean EHC	0.00831 (±0.010047)	5.27e-05 (±0.0000626)	0.00457 (±0.005837)	0.013711 (±0.01751)
	p-value	0.04365 *	0.04401 *	0.0215 *	0.0215 *
/o/	*Mean PPD*	0.005128 (±0.002545)	3.47e-05 (±0.0000197)	2.64e-03 (±0.001265)	0.007921 (±0.003796)
	Mean EHC	0.005525 (±0.004662)	3.75e-05 (±0.0000317)	0.002925 (±0.002681)	0.008774 (±0.008042)
	p-value	0.2467	5.31e-01	0.1479	0.1479

[3] A complete description of these features can be found in the Appendix of Toye, Kompalli (2021).
[4] https://github.com/aeesha-T/parkinsons_prediction_using_speech.

		Local Jitter	*Local Absolute Jitter*	*Rap Jitter*	*Ddp Jitter*
/u/	*Mean PPD*	0.003602 (±0.002028)	2.30e-05 (±0.0000137)	1.82e-03 (±0.00122)	0.005458 (±0.00366)
	Mean EHC	0.007342 (±0.008565)	4.71e-05 (±0.0000571)	4.71e-05 (±0.005042)	0.011704 (±0.015125)
	p-value	0.01048 *	0.02357 *	0.005963 **	0.005963 **

Table 12 - *Shimmer measures on vowel executions: means, st. deviations and statistical comparison (p-values) between PPDs and EHCs*

		Local Shimmer	*Local db Shimmer*	*apq3 Shimmer*	*apq5 Shimmer*	*Dda Shimmer*
/a/	*Mean PPD*	0.027403 (±0.021298)	0.25964 (±0.214383)	0.013286 (±0.010428)	0.015221 (±0.011861)	0.039859 (±0.031283)
	Mean EHC	0.072828 (±0.04221)	0.671414 (±0.371165)	0.037256 (±0.024488)	0.044148 (±0.028025)	0.111767 (±0.073464)
	p-value	2.2e-16 ****	2.2e-16 ****	3.594e-15 ****	2.20e-16 ****	3.59e-15 ****
/e/	*Mean PPD*	0.029551 (±0.022644)	0.282237 (±0.231909)	0.013932 (±0.010352)	0.015913 (±0.011252)	0.041796 (±0.031055)
	Mean EHC	0.059624 (±0.046414)	0.551995 (±0.397508)	0.029488 (±0.025365)	0.03815 (±0.037496)	0.088465 (±0.076095)
	p-value	1.40e-06 ****	4.48e-07 ****	5.66e-06 ****	3.58e-07 ****	5.66e-06 ****
/i/	*Mean PPD*	0.019129 (±0.013484)	0.18677 (±0.156311)	0.009133 (±0.007239)	0.010786 (±0.007944)	0.027398 (±0.021716)
	Mean EHC	0.052676 (±0.048381)	0.510718 (±0.441095)	0.025479 (±0.025769)	0.032395 (±0.035177)	0.076437 (±0.077308)
	p-value	2.74e-10 ****	1.65e-10 ****	9.60e-08 ****	1.366e-09 ****	9.60e-08 ****
/o/	*Mean PPD*	0.03095 (±0.01932)	0.289764 (±0.194466)	0.015355 (±0.009475)	0.017472 (±0.010994)	0.046065 (±0.028426)
	Mean EHC	0.058774 (±0.046212)	0.544526 (±0.412815)	0.027491 (±0.023018)	0.036019 (±0.033853)	0.082473 (±0.069054)
	p-value	6.83e-05 ****	4.23e-05 ****	0.001512 **	6.25e-05 ****	0.001512 *
/u/	*Mean PPD*	0.021482 (±0.019223)	0.201214 (±0.180082)	0.010756 (±0.010824)	0.012315 (±0.012971)	0.012315 (±0.032473)
	Mean EHC	0.048298 (±0.04619)	0.464424 (±0.418607)	0.022219 (±0.02524)	0.029579 (±0.034524)	0.066658 (±0.075719)
	p-value	1.71e-07 ****	1.26e-07 ****	6.26e-06 ****	3.01e-07 ****	6.26e-06 ****

Table 13 - *Harmonic-to-Noise Ratio on vowel executions: means, st. deviations and statistical comparison (p-values) between PPDs and EHCs*

	HNR		
	Mean PPD	*Mean EHC*	*p-value*
/a/	25.05944 (±8.191305)	16.28696 (±6.634543)	6.29e-13****
/e/	22.99675 (±7.757845)	18.7804 (±7.405272)	0.0003349***
/i/	27.49632 (±7.136836)	20.16619 (±7.762003)	1.04e-06****
/o/	23.8211 (±8.654639)	21.28797 (±7.730056)	0.03745*
/u/	30.44595 (±6.649103)	23.60019 (±9.587023)	5.22e-05****

4. *Discussion*

The results of our research are generally consistent with previous literature on parkinsonian speech: while dysphonia is commonly identified as one of the earliest signs of hypokinetic dysarthria (see § 3.3), our analysis also highlights spectral measures as some of the most significantly altered features. More specifically, patients in our sample showed lower means and higher variability on the Higuchi Fractal Dimension, as well as significantly greater variability in the central spectroid measure.

In addition to these metrics, the data gathered from Toye and Kompalli (2021) allowed us to extend the analysis to jitter, shimmer, and HNR: results on jitter variants were significant on the three corner vowels (/a/, /i/, /u/), while shimmer and HNR demonstrated high statistical significance across all vowel executions (/a/, /e/, /i/, /o/, /u/). Surprisingly, however, the variation in these metrics did not follow the expected patterns, as HNR yielded a higher value in Parkinson's patients, whereas the measures of spectral variability were greater in control subjects. This result, which we judge as noteworthy, prompts the need to conduct further studies on larger samples and with more attention to consistent data collection procedures.

Regarding speech segment articulation, our work confirms previous findings from foreign literature on vowel articulation indices and supports the first measurements conducted on an Italian-speaking population by Mirarchi *et al.* (2017), thereby reinforcing the validity of these metrics for the Italian language. In fact, the Vowel Articulation Index (VAI) decreases in pathological subjects of our sample, while FCR increases and the tVSA appears to be reduced, to such an extent that the difference is graphically visible in Fig. 2; another insight that can be drawn from the graph concerns the significantly lower values in PD patients' F1, likely due to reduced mobility of the jaw. Further analysis on biological sex, moreover, reveals a potential gender effect: vowel production appears especially impaired in male patients, while almost no relevant differences were shown by the comparison between female patients and their respective controls. It should be noted that the structure of the corpus at our disposal enabled the collection of formant frequencies on pure vowel execution tasks, eliminating other complexity factors related to rhythmic structure (e.g., the position of the syllable within wider prosodic units) or coarticulation:

positive results on this simpler execution mode, in which speakers could exclusively focus on good vowel articulation, can thus be considered further proof of the relevance of VAI, FCR, and tVSA as acoustic biomarkers. In general, based on the observation of vowels, we can conclude that the study of hypoarticulation of speech segments yields promising signals: further investigation in the future could include the extension of the analysis to consonant production, following the example of Antolík, Fougeron (2013).

Concerning prosody, our findings align with previous literature on pitch variability, indicating a significant reduction in patients of our sample. This reduction contributes to the perceived monotony which is commonly reported as a feature of parkinsonian speech.

As for duration measures, the difficulties of identifying consistent patterns of variation have already been discussed in § 1: rather than a uniform strategy, patients seem to adopt a range of different compensatory methods, including reduced speech rate, oral festination, or alterations in the distribution of pauses. In our corpus, these discrepancies appear to arise with the introduction of biological sex as an independent variable: in fact, different trends are found for male and female PPDs when compared to respective controls. For female patients, an increase in speech segment duration (SPEMEAN and SPEMEDIAN) is detected in all tasks, while the average total duration of silences (SILMEAN, SILMEDIAN) is reduced; as a result, a significant rise in Transformed Phonation Rate (TPR) and Standardised Pause Rate (measures inversely correlated with pauses) is observed, as well as a decrease in SPT. Results from male PPDs, on the other hand, vary in the opposite direction, with an increase in pause-related measures (SILMEAN and SILMEDIAN, along with SILSD) at the expense of speech segment durations (SPEMEAN and SPEMEDIAN), resulting in a decrease in TPR and VR. The overall interpretation of these data suggests that female patients exhibit a less disrupted speech pattern, characterised by longer speech segments and fewer pauses, while male patients would tend to produce shorter rushes of speech punctuated by long pauses. Interestingly, a similar result can be found in Skodda, Visser & Schlegel (2011b), who reported a highly significant reduction of pause durations in female patients; however, to our knowledge, gender effects on dysprosody have not been investigated before on Italian speakers.

Lastly, an observation can be made about the evaluation of UPDRS subgroups. As already stated in § 3.2, the comparison between the three subgroups was only possible for the first reading of the text since patients with more severe symptoms did not manage to execute the subsequent tasks. The difficulty in completing the protocol is – in itself – proof of the association between motor symptoms and speech impairment; this observation is further supported by data from the first reading of the text, as for many of the metrics (SILMEAN, SILSD, SILMEDIAN, SPEMEAN, SPESD, SPEMEDIAN, VR, TPR, SPR, HFractDM) the analysis clearly distinguishes patients in the most advanced stage from the two previous groups. In this regard, it is also worth mentioning that TRVSD (Temporal regularity of voiced segments) increases across the subgroups, making the comparison between

early- and later-stage patients significant: this result could be interpreted as consistent with previous studies on the rhythmic features of PD dysarthria, which have demonstrated an increase in %V and more regularity in patients' syllable durations.

5. *Conclusions*

Overall, the study successfully analysed various aspects of parkinsonian speech, contributing to the flourishing body of Italian research in the field. Our three-staged analysis revealed several differences between pathological subjects and controls: some of our findings, such as the results on pitch and vowel metrics, fully align with previous literature, while some others (outcomes on jitter, shimmer, and HNR) contradict the expectations and require further investigation. Furthermore, some novel elements have been introduced, notably a focus on spectral measures related to voice quality and the inclusion of gender effects on prosody, which, to the best of our knowledge, had not been previously investigated in the Italian language context.

Several research paths remain unexplored and could be addressed in future works. First of all, while the current study is mainly focused on quantitative methods, we deem it useful to extend the analysis to qualitative approaches which could eventually involve various features, ranging from consonant production to prosody and intonation. A second improvement to our work would consist in the expansion of the corpus, by including a larger number of speakers and additional tasks (e.g., spontaneous speech). Computational methods could also be implemented, by using the most significant measures as input features for classification algorithms: achieving promising outcomes in this domain would certainly pave the way for the development of automatic tools in support of the diagnosis. As a final point, among other variables that could possibly be explored, it is noteworthy to mention the response to surgical procedures such as Deep Brain Stimulation. As far as we know, there is a lack of Italian studies on this subject, while conflicting results can be inferred from foreign literature – including a worsening in speech impairment. We believe that further research (even on limited cohorts) is needed on this topic, as it could offer important arguments for or against this kind of treatment.

References

Antolík, T., Fougeron, C. (2013). Consonant distortions in dysarthria due to Parkinson's disease, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia. In *Interspeech 2013*, Lyon, France, 25-29 August 2013, 2152-2156.

Calzà, L., Gagliardi, G., Rossini Favretti, R. & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying Mild Cognitive Impairment and dementia. In *Computer Speech & Language*, 65, 101–113. https://doi.org/10.1016/j.csl.2020.101113

Cohen, H. (2003). Disorders of speech and language in Parkinson's disease. In Bédard, M.A., Agid, Y., Korczyn, A., Lespérance, P. & Chouinard, S. (Eds.), *Mental and*

Behavioral Dysfunction in Movement Disorders. New York, NY: Humana Press, 125–134. https://doi.org/10.1385/1-59259-326-7

Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D. & Girardi, F. (2017). "Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System". In *IEEE Access*, 5, 22199-22208. https://doi.org/10.1109/ACCESS.2017.2762475

Dworkin, J.P., Aronson, A.E. (1986). Tongue strength and alternate motion rates in normal and dysarthric subjects. In *Journal of communication disorders*, 19(2), 115–132. https://doi.org/10.1016/0021-9924(86)90015-8

Gagliardi, G. (2023). Natural language processing techniques for studying language in pathological ageing: A scoping review. In *International Journal of Language & Communication Disorders*, 59, 110-122. https://doi.org/10.1111/1460-6984.12870

Gagliardi, G., Tamburini, F. (2021). Linguistic biomarkers for the detection of Mild Cognitive Impairment. In *Lingue e Linguaggio*, XX, 3–31.

Gagliardi, G., Tamburini F. (2022). The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France, 21-23 June 2022, 5234-5242.

Gili Fivela, B., D'Apolito, S. & Di Prizio, G. (2020a). Labialization and Prosodic Modulation in Italian Dysarthric Speech by Parkinsonian Speakers: A Preliminary Investigation. In *Speech Prosody 2020*, 828-832. https://doi.org/10.21437/SpeechProsody.2020-169

Gili Fivela, B., D'Apolito, S. & Di Prizio, G. (2020b). Controllo motorio e disartria nella malattia di Parkinson: uno studio pilota sulla labializzazione. In *Studi AISV, 7*, 429-440. https://doi.org/10.17469/O2107AISV0000207

Gili Fivela, B., Iraci, M.M., Grimaldi, M., Zmarich, C. (2015). Consonanti scempie e geminate nel morbo di Parkinson: la produzione di bilabiali. In Vayra, M., Avesani, C., Tamburini, F. (Eds.), *Il farsi e il disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio*, 289-312. https://doi.org/10.17469/O2101AISV000018

Gili Fivela, B., Zmarich, C. (2005). Italian Geminates under Speech Rate and Focalization Changes: Kinematic, Acoustic, and Perception Data. In *InterSpeech 2005*, Lisbon, Portugal, 4-8 September 2005, 2897-2900.

Hanson, D.G., Gerratt, B.R. & Ward, P.H. (1984).Cinegraphic observations of laryngeal function in Parkinson's disease. In *The Laryngoscope*, 94(3), 348–353.https://doi.org/10.1288/00005537-198403000-00011

Hoehn, M.M., Yahr, M.D. (1967). Parkinsonism: onset, progression and mortality. In *Neurology*, 17(5), 427–442. https://doi.org/10.1212/wnl.17.5.427

Hunker, C.J., Abbs, J.H. (1990). Uniform frequency of parkinsonian resting tremor in the lips, jaw, tongue, and index finger. In *Movement disorders: official journal of the Movement Disorder Society,* 5(1), 71–77. https://doi.org/10.1002/mds.870050117

Liss, J.M., White, L., Mattys, S.L., Lansford, K., Lotto, A.J., Spitzer, S.M. & Caviness, J.N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. In *Journal of speech, language, and hearing research: JSLHR*, 52(5), 1334–1352. https://doi.org/10.1044/1092- 4388(2009/08-0208)

Lowit, A., Marchetti, A., Corson, S. & Kuschmann, A. (2018). Rhythmic performance in hypokinetic dysarthria: Relationship between reading, spontaneous speech and diadochokinetic tasks. In *Journal of communication disorders*, 72, 26–39. https://doi.org/10.1016/j.jcomdis.2018.02.005

Maffia, M., De Micco, R., Pettorino, M., Siciliano, M., Tessitore, A. & De Meo, A. (2021). Speech Rhythm Variation in Early-Stage Parkinson's Disease: A Study on Different Speaking Tasks. In *Frontiers in psychology*, 12:668291, 1-10. https://doi.org/10.3389/fpsyg.2021.668291

Maffia, M., De Micco, R., Tessitore, A. & De Meo, A. (2020). Acoustic characteristics of Italian Parkinsonian speech: a study on early-stage patients. In Botinis, A. (Eds.), *Proceedings of the 11th International Conference of Experimental Linguistics – Exling2020,* Athens, Greece, 12-14 October 2020, 117-120. https://doi.org/10.36505/ExLing-2020/11/0029/000444

Mirarchi, D., Vizza, P., Tradigo, G., Lombardo, N., Arabia, P. & Veltri, G. (2017). Signal Analysis for Voice Evaluation in Parkinson's Disease. In *2017 IEEE International 101 Conference on Healthcare Informatics (ICHI)*, Park City, UT, USA, 23-26 August 2017, 530-535. https://doi.org/10.1109/ICHI.2017.72

Moreau, C., Ozsancak, C., Blatt, J.L., Derambure, P., Destee, A. & Defebvre, L. (2007). Oral festination in Parkinson's disease: biomechanical analysis and correlation with festination and freezing of gait. In *Movement disorders: official journal of the Movement Disorder Society,* 22(10), 1503–1506.https://doi.org/10.1002/mds.21549

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease (2003). The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. In *Movement disorders: official journal of the Movement Disorder Society*, 18(7), 738–750. https://doi.org/10.1002/mds.10473

Pettorino, M., Busà, M.G. & Pellegrino, E. (2016). Speech rhythm in Parkinson's disease: a study on Italian. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, 8-12 September 2016, 1958–1961. https://doi.org/10.21437/Interspeech.2016-74

Pinto, S., Ghio, A., Teston, B. & Viallet, F. (2010). La dysarthrie au cours de la maladie de Parkinson. Histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie. In *Revue neurologique*, 166(10), 800-810. https://doi.org/10.1016/j.neurol.2010.07.005

Robertson, L.T., Hammerstad, J.P. (1996). Jaw movement dysfunction related to Parkinson's disease and partially modified by levodopa. In *Journal of neurology, neurosurgery, and psychiatry*, 60(1), 41–50.https://doi.org/10.1136/jnnp.60.1.41

Rusz, J., Cmejla, R., Ruzickova, H. & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. In *The Journal of the Acoustical Society of America*, 129(1), 350–367. https://doi.org/10.1121/1.3514381

Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J. & Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. In *The Journal of the Acoustical Society of America,* 134(3), 2171–2181. https://doi.org/10.1121/1.4816541

Rusz, J., Cmejla, R., Ruzickova, H. & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated. Parkinson's disease. In *The Journal of the Acoustical Society of America*, 129(1), 350–367. https://doi.org/10.1121/1.3514381

Sapir, S., Ramig, L.O., Spielman, J.L. & Fox, C. (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. In *Journal of speech, language, and hearing research: JSLHR,* 53(1), 114–125. https://doi.org/10.1044/1092- 4388(2009/08-0184)

Skodda, S., Grönheit, W. & Schlegel, U. (2012). Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. In *PloS one*, 7(2): e32132, 1-8. https://doi.org/10.1371/journal.pone.0032132

Skodda, S., Visser, W. & Schlegel, U. (2011a). Vowel articulation in Parkinson's disease. In *Journal of voice: official journal of the Voice Foundation*, 25(4), 467–472. https://doi.org/10.1016/j.jvoice.2010.01.009

Skodda, S., Visser, W. & Schlegel, U.(2011b). Gender-related patterns of dysprosody in Parkinson disease and correlation between speech variables and motor symptoms. In *J. Voice 25* (2011b) 76–82. https://doi.org/10.1016/j.jvoice.2009.07.005

Toye, A.A., Kompalli, S. (2021). Comparative Study of Speech Analysis Methods to Predict Parkinson's Disease. In *ArXiv*.:2111.10207, 1-10. https://doi.org/10.48550/arXiv.2111.10207

DOI: 10.17469/O2111AISV000006

BARBARA GILI FIVELA, RAFFAELLO CASERTA, SONIA D'APOLITO,
ANNA CHIARA PAGLIARO, VINCENZO SALLUSTIO, MARCO SINISCALCHI

Phonetic analysis and deep neural network examination for dysarthria detection beyond sociolinguistic differences: a comparison

The goal of the paper is to compare a phonetic and an automatic evaluation, via Convolutional Neural Networks, of both dysarthric and control speech. On the one hand, we want to observe if and how phonetic and automatic classifications of dysarthric and healthy speech perform differently, given their specificities; on the other hand, we aim at checking if the automatic analysis is deceived by sociophonetic traits, taking them as instances of non-precise speech production. For the purpose of the analysis, we used a corpus of acoustic recordings of speech produced by 15 male parkinsonian dysarthric speakers (8 from Bari; 7 from Lecce) and 10 male control speakers from the same areas. Part of the corpus was used for fine-tuning the Neural Network that was previously trained on Spanish data; the remaining part of the corpus was used for testing the automatic and the phonetic evaluation. The comparison shows that the automatic and phonetic analyses seem to offer similar results in terms of classification, also with respect to sociolinguistic features. However, they could be successfully integrated as they suggest different information on the speakers and their speech.

Keywords: Tranfer learning, Italian, Spanish, individual strategies, sociophonetics.

1. *Introduction*

Speech accuracy is considered a good metric for assessing the severity level of motor speech disorders, such as dysarthria. Among the objective methods available to assess accuracy, phonetic analysis provides very detailed information, shedding light on individual strategies, even with reference to specific speech sounds or gestures. For instance, acoustic studies have shown that vowels produced by parkinsonian dysarthric speakers (PDs) occupy a reduced acoustic space compared to those of healthy controls (HCs), and show specific acoustic characteristics, such as increases in jitter and noise-to-harmonics ratio (a.o., Bang, Min, Sohn & Cho, 2013; Tjaden, 2000); as for consonants, they may be less precisely produced (a.o. Kim, Gurevich, 2021). Phonetic investigation results may be correlated to severity levels that are actually guaranteed by overall, usually subjective evaluations of dysarthria (e.g., Robertson's Dysarthria Profile; Robertson, 1982; Fussi, Cantagallo, 1999) or that may offer an overall and not speech-specific or dysarthria-specific evaluation (e.g., Unified Parkinson's Disease Rating Scale – UPDRS – and Hoehn and Yahr; Skodda, Gronheit & Schlegell, 2012; Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, 2003). Nevertheless, in the case of phonetic

investigations, especially in the case of articulatory studies, analysed speech sounds or even populations may often not be wide enough to allow for robust generalization (Wong, Murdoch & Whelan, 2010; Wong, Murdoch & Whelan, 2011).

On the other hand, automatic learning algorithms are also used for objectively evaluating speech accuracy by considering global speech characteristics, even though no detailed observation on specific sound properties is provided (Wang, Fu, Shao, Chang, Ren, Chen & Ling, 2022, but see also Audibert, Fougeron, Fredouille, Meunier & Panseri, 2010). From a technical standpoint, the speech impairments exhibited by individuals with Parkinson's Disease (PD) have been analyzed by computing features associated with four dimensions: phonation, articulation, prosody, and intelligibility. Despite the established effectiveness of traditional feature extraction methods, recent years have witnessed the successful application of deep learning techniques for evaluating particular aspects of speech, such as the detection and monitoring of PD. These diverse speech analysis techniques encompass the utilization of technologies ranging from Support Vector Machines (SVM) to Convolutional Neural Networks (CNN). According to Vásquez Correa, Arias, Orozco-Arroyave & Nöth (2018), it is possible to achieve automatic recognition of speech characteristics by employing a CNN to analyze features extracted directly from the speech audio and corresponding to distinctive articulatory segments. This approach allows for the automated assessment of dysarthria severity through the examination of specific articulatory patterns realizing speech segments. However, it is noteworthy that the exploration of entire words or phrases remains an unexplored area in this context. Further research is required to delve into the potential insights that a comprehensive analysis of phonation and articulation across complete linguistic units could provide in understanding and assessing PD-related speech impairments.

A further issue regards sociophonetic features, which characterize everyone's speech since they are the manifestation of each speaker sociolinguistic identity, being due to various factors including the speech characteristics of the area from which the speaker comes from. Some sociolinguistic features may be the result of reduction or lenition phenomena that have taken place during the evolution of a specific dialect and may have an impact on the variety of the language spoken in the same area. Crucially, phenomena such as the reduction of vowels and the lenition of consonants may also be considered as the outcome of a diminished accuracy and precision in speech articulation in pathological speakers. In phonetic or subjective evaluation, they may be recognized as such, thanks to a phonology-driven approach to the phonetic analysis or to a sort of normalization made by speech therapists; but how do automatic analyses work in these cases?

Questions then arise concerning 1) the possible differences in efficacy and accuracy in distinguishing healthy and pathological speech by analyzing specific speech events, as in most of phonetic studies, or by exploiting automatic evaluations which consider the acoustic signal as a whole. In other words, one may wonder if the in-depth phonetic analysis of specific phenomena is representative of the overall speech accuracy or it rather describes too local events that do not necessarily fit

the severity level assigned to the experimental subject or to an entire sentence or item s/he produces; on the other hand, one may also wonder if 2) automatic analyses of speech, blind to very specific linguistic knowledge, may be deceived by sociophonetic traits, taking them as instances of non-precise speech production.

In this paper, we try to answer these two questions by comparing a phonetic (acoustic) analysis and an automatic evaluation of both dysarthric and control speech. This work is part of a larger PRIN project aiming at exploiting phonetic analysis of dysarthric speech by speakers of different varieties of Italian to develop clinical tools that face sociophonetic variation. Phonetic analyses are carried out semiautomatically (via scripts) and statistically, while automatic evaluations are carried out by means of CNNs (plus a check by means of SVMs), implemented on purpose for the project mentioned above.

The article is structured as follows: the state-of-the-art for investigations on dysarthric and control speech, with specific attention to the varieties considered here, is offered in § 2, and that for investigations based on deep neural networks for dysarthria detection is given in § 3; § 4 reports goals and hypotheses; the methodology is described in § 5, and the results of the present investigation are reported in § 6. The discussion and conclusions found in § 7 close the paper.

2. *Dysarthric and control speech in Lecce and Bari Italian*

As mentioned in § 1, in speech produced by PDs affected by (hypokinetic) dysarthria consonants are articulated with reduced accuracy and vowels are reported to occupy a reduced acoustic space compared to those of HCs, so that the vowel space could be a possible biomarker for the early detection of dysarthria (Roland, Huet, Harmegnies, Piccaluga, Verhaegen & Delvaux, 2023).

As far as Italian is concerned, experiments on acoustic and articulatory vowel space have shown that reduction takes place, but it is not homogenous (Gili Fivela, d'Apolito & Sigona 2022a). In particular, asymmetries in reducing the acoustic vowel space may be found, in line with articulatory data showing that some subjects appear to compensate in order to get a specific acoustic output. Concerning consonants, in line with results on other languages, the closure phase in plosives has been found to be not properly realized or, in any case, not articulated as HC do. However, plosives as well as fricatives are produced as accurately as possible as they are relevant phonemes in the languages. On the other hand, a sociolinguistic marker such as aspiration may be not preserved through compensatory strategies as much as the difference between phonologically relevant segments (Gili Fivela, d'Apolito & Pagliaro, 2022b). However, the disease seems to have a stronger impact on the production of the plosives than that of fricatives (Gili Fivela, Pagliaro & d'Apolito, 2023), even though correlates such as duration may change in both plosives and fricatives compared to HC segments (Gili Fivela et al., 2022b).

Crucially, varieties of Italian show different sociophonetic features that may be interestingly consider in the investigation of dysarthria. For instance, the varieties

of Italian spoken in Bari and Lecce (Southern Italy), belong to different dialectal isoglosses (South vs. Extreme South; Rohlfs, 1949) and are characterized by different sociophonetic markers, such as unvoiced plosive strong aspiration in Lecce, and vowel reduction or deletion in pre- or post-tonic position in Bari (where the system is eptavocalic, while it is pentavocalic in Lecce; Loporcaro, 1997; Romano, Manco, 2004). Thus, some of the characteristics variably distributed in the two varieties go in the direction expected in dysarthric speech, that are a variable phasing of laryngeal and supralaryngeal gestures (Best, 1995; Browman, Goldstein 1986), such as in unvoiced plosive aspiration in the Lecce variety, and a tendency to reduce vowels, such as in vowel reduction in unstress position in Bari Italian. Some research on the impact of sociolinguistic features on the dysarthria progression has been mentioned above and the issue is considered also in the analysis presented in the next sections.

3. *Deep neural networks for dysarthria detection*

The ultimate goal in using deep neural networks is the development of a diagnostic tool or diagnostic support for clinical personnel in the early diagnosis of PD through the automatic analysis of patients' speech and the identification of dysarthria symptoms. The system can provide an estimate of severity at the network's output, or it can simply binary classify characteristic features extracted from a suitable dataset. The desired diagnostic tool focuses precisely on analysing a whole recording, that is, a set of selected word or sentences. The objective is to prompt a patient to undergo an analysis of her/his speech under conditions compatible with a hospital setting, rather than in a laboratory specifically designed for the analysis of audio recordings.

Convolutional Neural Networks have proven to be among the most effective models for impaired speech analysis, as demonstrated by Vásquez Correa et al. (2018). CNNs belong to a class of deep neural networks designed to process grid-like data structures, such as images and, in the context of speech, spectrograms generated from audio signals. Their architecture features convolutional layers capable of automatically learning hierarchical representations, effectively capturing both local and global patterns. This capability makes CNNs especially well-suited for speech analysis tasks, where preserving the spatial relationships within input data is essential.

Despite their strengths, CNNs present certain challenges in speech analysis applications. Chief among these is the requirement for large volumes of labeled data for effective training, which can be both time-consuming and resource-intensive to obtain. Furthermore, transferability across datasets poses a concern, as CNNs trained on one dataset may not perform well on another related task. To address this, transfer learning — where a pre-trained model is fine-tuned on a new but related task — can be employed to leverage existing knowledge and improve generalization to new speech analysis domains.

4. *Aims and hypotheses*

The main goal is to observe if and how phonetic and automatic classifications of dysarthric and healthy speech perform differently, given their obvious specificities. A second goal is to check if automatic analyses of speech may be deceived by sociophonetic traits, taking them as instances of non-precise speech production.

The main hypothesis is that the phonetic and automatic procedures may classify speech consistently but do not allow for easy comparison, given the specificities mentioned in § 1. Still, they may be integrated, offering equally valuable, though somehow different, evaluations of the same speech. Further, another hypothesis is that sociophonetic features may be investigated by phonetic analyses, but they may be misleading in automatic ones.

5. *Methodology*

5.1 General methodology

A corpus was collected in order to reach the goals and check the hypotheses stated in § 4. The overall corpus considered here includes acoustic recordings of speech produced by 15 PD speakers (males, 8 from Bari and 7 from Lecce) and 10, though effectively 9[1], HC speakers (males, 5 from Bari and 4 from Lecce).

Participants were selected so as to be highly comparable in terms of age (PD: avg. 62,82, SD 8,17; HC: avg. 60,87, SD 8,75) and cognitive capabilities (Montreal Cognitive Assessment: PD=avg. 25, S.D. 1,1; HC= avg. 26,57, S.D. 1,61). HC reported no previous motor disorders and, consistently, showed a normal motor profile (Therapy Outcome Measure for Dysarthria – TOM-Dys: avg. 5, SD 0; Robertson's dysarthria Profile: > 134). PD participants varied greatly as for the number of years from diagnosis (avg.10,67, SD 6,62), but showed less variation as for their motor profile (TOM-Dys: avg. 3,71, SD 0,49; Robertson's Dysarthria Profile: <127).

Speakers were recorded reading aloud sentences included in 3 different short passages (14 sentences in total), sentences presented in isolation (9, produced with no specific expectation and attitude), and sentences preceded by a context that induced a specific, intended interpretation (16, including various modalities, such as declaratives and interrogatives, that are yes-no and wh-questions). Three repetitions were collected during recordings, which took place in a quiet room of the hospitals in Lecce and Bari.

The total amount of recorded speech considered here (excluding all pauses) is 56:75 minutes, 20:03 of PD dysarthric speech and 36:73 of HC speech.

Part of the corpus described above was used for fine-tuning and part for testing the automatic and phonetic evaluation. As for the former, the speech produced by 9 PD (5 from Bari and 4 from Lecce) and by 4 HC (2 from Bari and 2 from Lecce)

[1] Speech materials produced by one HC from Lecce were excluded due to technical issues.

was used; for the testing phase, the speech recorded by the remaining speakers was adopted, that is 6 PD speakers (3 from Bari and 3 from Lecce) and 5 HC (3 from Bari and 2 from Lecce). Speakers were assigned to the training or test set so as to have representative speakers in both sets and to have two groups as balanced as possible.

5.2 Methods for automatic analyses

In our study, we explored multiple CNN configurations. The configuration that achieved the best performance comprised four identical blocks, each containing a 2D convolutional layer, followed by batch normalization and an activation function. A spatial dropout layer was inserted between the first two layers to mitigate overfitting. Another configuration tested involved two blocks, each consisting of two stacked 2D convolutional layers, followed by ReLU activation, pooling, and dropout. Various hyperparameter adjustments were applied to optimize classification performance[2].

The automatic system was developed using a CNN trained in a cross-linguistic framework. Initially, the CNN was pre-trained on the PC-GITA database (Orozco-Arroyave, Arias-Londoño, Vargas-Bonilla, González-Rátiva & Nöth, 2014), a Spanish-language speech corpus designed for the analysis of individuals with PD. This dataset includes recordings from 50 PD patients and 50 HC speakers.

Following this pre-training phase, the model was fine-tuned using speech data from Italian speakers, which were excluded from the testing set to ensure unbiased evaluation. Due to the high level of noise in the Italian recordings, a preprocessing step was implemented to enhance audio quality. This involved the use of a speech enhancement system based on a Deep Neural Network (DNN), as proposed by Siniscalchi (2021), to improve signal clarity prior to PD detection.

The system was trained to perform a binary classification task, predicting the HC (0) or the PD (1) status of speakers who realized specific speech samples.

5.3 Methods for phonetic analyses

The analysis was carried out by means of a semi-automatic procedure. First, forced alignment was realized via MAUS (Schiel, 1999) trained on Italian, and then manual correction of segment boundaries and labelling of stop burst was performed in PRAAT (Boersma, 1992-2022). PRAAT scripts were then used to measure segment and Voice Onset Time (VOT) duration, as well as formant values (F1, F2), measured as the average value over the central 30 ms of vowels, for both vowels in stress and unstress syllables.

[2] For comparison, Support Vector Machines (SVMs) were also evaluated for classifying dysarthria severity. In our experiments, we employed standard SVM structures and fine-tuned the hyperparameters *C* and *gamma* within a range defined by powers of ten. Training was carried out using both linear and radial basis function (RBF) kernels to assess performance under different conditions. Related results are not discussed in this paper.

The analysis focussed on aspects that could highlight differences in accuracy in speech production and in the realization of sociophonetic traits. As for the accuracy in production, the analysis concerned the F1xF2 vowel space used for vowels /i/, /a/, /u/ in stress position, and the Vowel Articulation Index (VAI[3] – Roy, Nissen, Dromey & Sapir, 2009; Sapir, Ramig, Spielman & Fox, 2010; Skodda, Visser & Schlegel, 2011) to get information on the possible centralization of vowels, therefore again on the vowel space. Concerning consonants, the attention was focused on the accuracy in unvoiced stop production, analysed in terms of the number (percentage) of stops realized with a clear burst. Sociophonetic traits were analyzed by looking for differences across varieties and speakers both in the centralization of vowels /i/, /a/, /o/[4] in unstressed syllables (observed, again, by measuring the F1xF2 vowel space and the VAI), and in the aspiration of stops, by measuring differences in stop VOT duration (as a percentage of C duration).

Descriptive statistics, i.e. z-test of proportions, was used for analyzing percentages and T-test was used for analyzing VAI results.

6. *Results*

6.1 Automatic binary classification

Overall results obtained by means of the automatic evaluation in a binary classification task (HC vs. PD) are reported in the left part of Table 1, while accuracy results obtain for each subject are shown in the right part. Accuracy, sensitivity and specificity values range between 0.72 and 0.78, pointing to a satisfactory overall performance. Considering accuracy values for each subject, greater average accuracy is obtained for Bari PD and Lecce HC subjects. However, average values for Bari HCs and Lecce PDs are majorly affected by performances related to two subjects respectively (HC_BA_2 and PD_LE_8).

Table 1 - *Automatic evaluation in a binary classification task (HC vs. PD): Overall performance (left table) and by subject accuracy results (right table)*

Binary classification Overall	
Accuracy	0.758
precision	0.752
Recall	0.753
f1	0.752
roc_auc	0.753
sensitivity	0.784
specificity	0.722

Binary classification PD			*Binary classification HC*		
Speaker	Accuracy	Avg	Speaker	Accuracy	Avg
BA_3	1.000	0.96	BA_2	0.301	0.65
BA_5	0.907		BA_3	0.716	
BA_6	0.964		BA_5	0.927	
LE_4	0.860	0.66	LE_3	0.833	0.92
LE_6	0.836		LE_5	1.000	
LE_8	0.275				

[3] A lower VAI score points to a centralization of vowel formants.

[4] The vowel /o/ was chosen due to the lack of unstressed /u/.

The prediction of each participant belonging to the PD (1) or the HC (0) group was also calculated for each task (see Table 2). Results show that each speaker is quite accurately rated along a continuum ranging from HC to PD: on the one hand, PD subject productions received average evaluations greater that 0.6-0.7, besides speaker PD_LE_8 who was rather rated as a HC subject (0.2); on the other hand, rates associated to HC subjects are lower than 0.3-0.4, besides speaker HC_BA_2 who received an average evaluation around than 0.5.

As far as the task is concerned, besides PD_LE_8, PD productions seem to get worse predictions in sentences than in passages, while on the contrary HC productions in passages are always worse than those concerning sentences. Finally, in some cases, the rating of some subjects shows a relatively high difference depending on the task (SD around 0.2 for speakers PD_LE_4, HC_BA_3 and HC_LE_3).

Basically, PD productions were rated as HC in about 25.54 % of sentences, while HC realizations were rated as PD in about 19.14 % of sentences.

Table 2 - *Predictions of PD vs. HC status for each subject and each task*

		Sentences	*Passages*	*Average*	*SD*
PD	BA_3	1	1	*1*	0
	BA_5	1	1	*1*	0
	BA_6	*0.969*	0.929	*0.949*	0.028
	LE_4	0.634	*0.929*	*0.781*	0.208
	LE_6	*0.720*	0.643	*0.681*	0.055
	LE_8	*0.250*	0.143	*0.196*	0.076
HC	BA_2	0.515	*0.538*	*0.527*	0.016
	BA_3	0.208	*0.500*	*0.354*	0.206
	BA_5	0	*0.071*	*0.036*	0.051
	LE_3	0.125	*0.429*	*0.277*	0.215
	LE_5	0	0	*0*	0

As far as the impact of the variety is concerned, automatic rating of PD speech produced by Bari speakers tend to be slightly worse (closer to 1) than that produced by speakers from Lecce. On the other hand, the ratings of HC seem to be more balanced across the two varieties. In fact, excluding HC_BA_2 and PD_LE_8 who are particularly deviant and seem to correspond to borderline speakers, PDs in Bari and Lecce on average are rated 0,983 and 0,731, respectively; HCs in Bari and Lecce are rated 0,195 and 0,138 respectively.

6.2 Phonetic evaluation

The analysis of segmental characteristics in the whole set of test sentences rated automatically offers interesting results.

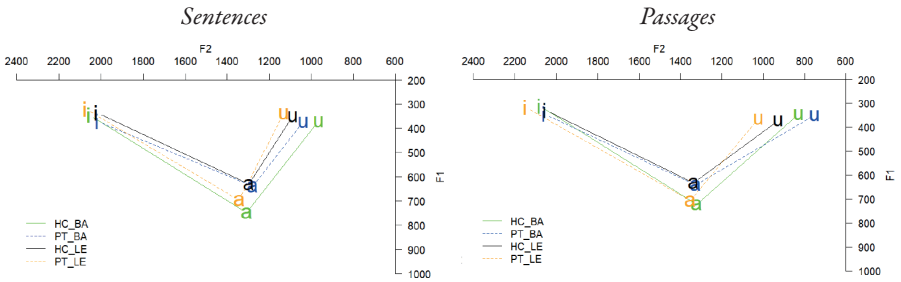
As for the production of stressed syllables, the VAI values show a slight centralization of vowels in PD productions in comparison to HC ones, though

differences are not significant either overall (Table 3, two upper rows) or when comparing speakers within variety (Table 3, bottom rows and Figure 1). However, plotting the raw Hz data, the vowel space seems to be reduced due to a decreased F1 value (reduced tongue depression) in the productions of /a/, at least by HC_LE and PD_BA speakers (Figure 1).

Table 3 - *Vowels/i/, /a/, /u/ in stressed syllable realized in sentences and passages by HCs and PDs from Bari and Lecce: VAI mean values and T-test (p-value)*

		Sentences	*T-test*	*Passages*	*T-test*
	HC	0,914	0,510	0,961	0.795
	PD	0,886		0.941	
BA	HC	0.948	0.242	1.0002	0.842
	PD	0.874		0.981	
LE	HC	0.862	0.494	0.901	0.989
	PD	0.898		0.899	

Figure 1 - *Vowels/i/, /a/, /u/ in stressed syllables realized in sentences (left) and passages (right) by HCs and PDs from Bari and Lecce (colors/lines): F1xF2 (Hertz) plots*

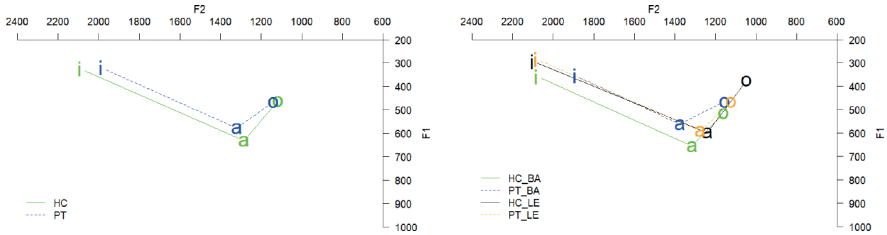


Concerning the realization of unstressed syllables, the VAI values show a slight centralization of vowels in PD productions in comparison to HC at least in sentences. The values do not change significantly (Table 4, two upper rows), besides a slightly significant difference for Bari speakers in sentences and Lecce speakers in passages (Table 4, bottom rows). Plotting the raw Hz data allows us to observe that the vowel space seems to be reduced due to a decreased F1 value (reduced tongue depression) in the productions of /a/ and a reduced F2 value (reduced advancement of the tongue) in /i/ (Figure 2, left); however, strategies may change depending on the variety, as the reduction may also concern the production of the /o/ vowel (e.g. Lecce PDs, Figure 2, right).

Table 4 - *Vowels/i/, /a/, /o/ in unstressed syllable realized in sentences and passages by HCs and PDs from Bari and Lecce: VAI mean values and T-test (p-value)*

		Sentences	*T-test*	*Passages*	*T-test*
	HC	0,851	0,113	0,783	0.624
	PD	0,784		0.794	
BA	HC	0.816	0.065	0.804	0.137
	PD	0.737		0.766	
LE	HC	0.904	0.285	0.751	0.058
	PD	0.831		0.822	

Figure 2 - *Vowels/i/, /a/, /o/ in unstressed syllables realized in sentences by all HCs and PDs (left) and by HCs and PDs from Bari and Lecce (right): F1xF2 (Hertz) plots*



The accuracy in plosive consonant production was checked by verifying the presence of a burst. Percentages of accurate productions are reported for HCs and PDs together with the results of z-test comparison (Table 5). Among the unvoiced plosives, PDs produced less accurately than HCs both bilabial stops (in both sentences and passages) and alveolar ones (only in sentence production), while no significant difference was found in the production of alveolars in passages and velars.

As far as the difference between Lecce and Bari speakers is concerned (see ** in Table 6), the only significant results concern the PD realization of /p/ in sentences, which is more accurate in Lecce PD speaker productions, and the /t/ realization in passages by HCs, which is again more accurate in the Lecce group.

Table 5 - */p/, /t/, /k/ production in sentences and passages by HCs and PDs: Accuracy of stop production as the percentage and total number of plosives and z-test comparison*

		Sentences	*z-test*	*Passages*	*z-test*
/p/	HC	100 %(320)	<0,0001	87,67 %(73)	<0.05
	PD	84,62 %(403)		69,32 %(88)	
/t/	HC	94,61 %(167)	<0,05	86,70 %(188)	n.s.
	PD	87,79 %(213)		84,75 %(236)	
/k/	HC	-	-	69,57 %(46)	n.s
	PD	-		65,22 %(69)	

Table 6 - *Accuracy in /p/, /t/, /k/ production in sentences and passages by HCs and PDs from Bari and Lecce (percentage and total number of plosives) and z-test comparison*

			Sentences	*z-test*	*Passages*	*z-test*
BA	/p/	HC	100 %(147)	<0,0001	92,86 %(42)	<0,05
		PD	74,4 %(168)**		68,18 %(44)	
	/t/	HC	93,75 %(80)	-	77,14 %(105)**	<0,05
		PD	88,64 %(88)		84,95 %(113)	
	/k/	HC	-	-	64 %(25)	-
		PD	-		72,5 %(40)	
LE	/p/	HC	100 %(173)	<0,0001	80,64 %(31)	-
		PD	91,91 %(235)**		70,45 %(44)	
	/t/	HC	95,40 %(87)	-	98,8 %(83)**	<0,05
		PD	87,2 %(125)		84,55 %(123)	
	/k/	HC	-	-	76,2 %(21)	-
		PD	-		55,17 %(29)	

7. *Discussion and conclusions*

The main goal of the work presented here was to observe if and how phonetic and automatic classification of dysarthric and healthy speech can differ on the same data. Further, a second aim was to check the role of sociophonetic difference in the speech data, in particular to understand if automatic analyses of speech may be deceived by sociophonetic traits that, when sociophonetics is not explicitly taken into account, could be considered as instances of inaccuracy in speech production. For the purpose of the analysis, a corpus was collected (56:75 minutes of articulated speech), including acoustic recordings of speech produced by 15 PD speakers (males, 8 from Bari and 7 from Lecce) and 10 HC speakers (males, 5 from Bari and 4 from Lecce, as one speaker was discarded for technical issue). Part of the corpus was used for fine-tuning the system for automatic analysis, and part for testing both the automatic and phonetic evaluation.

As for the overall performance of the automatic system, accuracy, sensitivity and specificity values range between 0.72 and 0.78, pointing to a satisfactory overall performance given the task and the data. Considering accuracy values for each subjects, greater average accuracy is obtained for Bari PD and Lecce HC subjects, with results that are clearly deviant in the case of two subjects, who are taken to be borderline (HC_BA_2 and PD_LE_8). The prediction of each participant belonging to the PD (1) or the HC (0) group turned out to rate speakers' production along a continuum ranging from HC to PD, even though results also depend on the task. In particular, PD productions seem to get worse predictions in sentences than in passages, while the opposite is true for HCs. The rating of subject productions along a continuum may easily correspond to different speaker performances; further, the different results concerning the task could correspond to an actual greater difficulty for PD in interpreting sentences than in reading

passages. In fact, sentences are shorter and clearly separated from each other, and the repetitive starts could possibly challenge PDs; further, in most cases required an interpretation that depended on the previous context. Finally, as far as the impact of the variety is concerned, the automatic rating of PD speech seems to be slightly worse in the case of Bari than Lecce speakers; however, the ratings of HCs seems to be more comparable across variety. The tendency may than be due to possible slight differences in the severity of dysarthria, as even though participants were chosen to represent a similar severity level (as measured by TOM-Dys and RDP; see § 5.1), Bari PDs could be slightly more compromised than the Lecce one. On the other hand, absolutely external factors, such as the quality of the recordings, seem to be not at play, as HC recordings took place in the same setting used for PDs and the result of their analysis seems to be comparable across varieties.

The phonetic analysis of vowels and stops produced in the all set of test sentences rated automatically was realized. As far as the production of vowels in stressed syllables is concerned, a slight centralization of vowels in PD production was found in comparison to HC realization, apparently mainly due to the reduced tongue lowering for /a/ (at least in Bari PDs and Lecce HCs), though PDs' VAI values were not significantly different from HCs' either overall or within variety. More interesting results in differentiating PDs and HCs are found in analyzing vowels realized in unstressed syllables: a slight centralization is observed for vowels in PD productions in comparison to HC at least in sentences for Bari speakers. However, the vowel space seems to be reduced not homogeneously, and the reduction does not necessarily involve the low vowel /a/ (e.g., the F1xF2 plots suggest that in sentence production, Lecce PDs seem to change the /o/ vowel aperture more than the /a/ vowel). As for the accuracy in unvoiced plosives, PDs produced less accurately than HCs both bilabial stops (in both sentences and passages) and alveolar ones (only in sentence production), while velars and alveolars in passages do not significantly differ. Finally, as for sociolinguistic differences the only significant results concern the PD realization of /p/ in sentences and that of /t/ by HCs in passages: in both cases, accuracy is greater in the Lecce group.

A comparison of the results of the automatic and the phonetic analysis aimed at distinguishing PD and HC subjects confirms our first hypothesis. The phonetic and automatic procedures may classify speech consistently, but do not allow for easy comparison: the automatic analysis rates speakers along a continuum ranging from HC to PD, even though results also depend on the task as PD productions seem to get worse predictions in sentences. On the other hand, the acoustic analysis focusing on specific segments and measures, does not consistently tease apart the two populations. Rather, it signals tendencies to reduce the vowel space, especially when vowels in unstressed syllables are considered, and produce unvoiced plosive consonants less accurately, especially bilabials and alveolars in sentences. Differences depending on the task are detected also in acoustics, as the results on alveolars show. Therefore, none of the two methods proved to be ideal. Rather, automatic and phonetic analysis could be successfully integrated depending on to goal of the

classification: if one also wants to analyze the reason for the results, phonetic analysis is to be preferred, whereas, if one wants to obtain a result very quickly, automatic analysis is, not surprisengly, to be preferred.

Regarding the second goal, that is understanding if automatic analyses of speech may be deceived by sociophonetic traits, results are not conclusive. The automatic rating of PD speech seems to be slightly worse in the case of Bari than Lecce speakers and, in theory, this could be related to differences in vowels (centralized in Bari, at least in unstress syllables) and consonant production (potentially considered as more accurate in Lecce due to the longer VOT expected for aspiration, requiring a clear burst). Consistently, also the phonetic analysis suggest that Lecce PD speakers are more accurate. However, a possible difference between the automatic and the phonetic approach may lie in the impact of the quality of the recordings: phonetic analyses, such as that presented here, are theory-driven and carried out semiautomatically, ensuring a systematic check of the data and speech phenomena, whereas a completely automatic treatment could be more easily affected by differences in the quality of recordings, that do not relate with speech production per se. Crucially, such difference does not mean that the automatic analysis was misled by sociophonetic traits. Further, if more data were available, the system could be trained on speech of a specific variety, making the system more robust as for sociophonetic features.

All in all, given the data considered here, automatic and phonetic analysis seem to offer similar results. However, they could be successfully integrated depending on to goal of analysis: if results are to be analyzed to get a better comprehension of the way speech is articulated, also taking into account sociophonetic features, the phonetic analysis is to be preferred; on the other hand, if a classification is to be achieved very quickly, with no specific interest in what speech characteristics are making the difference, the automatic analysis is, no surprise, to be preferred.

Acknowledgements

This work was funded by the PRIN 2017 project 2017JNKCYZ. We thank M.L. Fiorella and her collaborators, M. Bosco and D. Sciancalepore, for enrolling and recording experimental subjects in Bari; V. Sallustio, A. Trittola and R. Mitolo for subjects in Lecce. We warmly thank all the pathological and control subjects who participated in the experiment.

References

Audibert, N., Fougeron, C., Fredouille, C., Meunier, C. & Panseri, O. (2010). Evaluation d'un alignement automatique sur la parole dysarthrique. In *Mons: Actes des 28e Journées d'Etudes sur la Parole* (JEP'10), 353-356.

Bang, Y.-I., Min, K., Sohn, Y.H. & Cho, S.-R. (2013). Characteristics of vowel sounds in patients with Parkinson disease. In *NeuroRehabilitation*, 32(3), 649-54.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Best, C.T. (1995). A direct realist view of cross-language speech perception. In Strange, W. (a cura di), *Speech perception and linguistic experience: issues in cross-language research*. York Press, 171–204.

Browman, C.P., Goldstein, L. (1986). Articulatory gestures as phonological units. In *Phonology*, 6, 151-206.

Fussi, F., Cantagallo, A. (1999). *Profilo di valutazione della disartria*. Adattamento italiano del test di Robertson, raccolta di dati normativi e linee di trattamento, Ed. Omega.

Gili Fivela, B., d'Apolito, S. & Sigona, F. (2022a). Vowel space in hypokinetic dysartria: Preliminary investigations. In Dovetto, F.M., Raso, T. & Sorianello, P. (a cura di), *Le patologie del linguaggio: studi e risorse tra crossdisciplinarità e interdisciplinarità, Chimera (ISSN 2386-2629)*, 9, 147-164.

Gili Fivela, B., d'Apolito, S. & Pagliaro, A.C. (2022b). Tra economia dello sforzo e accuratezza nella disartria ipocinetica/ Between economy of effort and speech accuracy in hypokinetic dysartria. In Orrico, R., Schettino, L. (a cura di), *La posizione del parlante nell'interazione: atteggiamenti, intenzioni ed emozioni nella comunicazione verbale [The position of the speaker in interaction: attitudes, intentions, and emotions in verbal communication]*, *Studi AISV 9*. Milan: Officinaventuno, ISBN 978-88-97657-63-7, 99-114.

Gili Fivela, B., Pagliaro, A.C. & d'Apolito, S. (2023). Phonological and sociophonetic information in dysarthric speech: a first articulatory investigation on Italian. In Skarnitzl, R., Volín, J. (a cura di), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, 7-11 August 2023, Guarant International, 3937-3941.

Kim, H., Gurevich, N. (2021). Positional asymmetries in consonant production and intelligibility in dysarthric speech. In *Clinical Linguistics & Phonetics*, 37(2), 125-142.

Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease (2003). State of the Art Review The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations. *Movement Disorders*, Vol. 18(7), 738–750.

Loporcaro, M. (1997). Puglia & Salento. In Maiden, M., Parry, M. (a cura di), The Dialects of Italy, London: Routledge, 338-348.

Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C. & Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Language Resources and Evaluation Conference*, (LREC), 2014, 342–347.

Robertson, S.J. (1982). Dysarthria profile. Background and development. In *College Speech Therapist Bullettin*, 359.3.82.

Rohlfs, G. (1949). *Historische Grammatik der Italienischen Sprache und ihrer Mundarten*. Vol. 1. Lautlehre. Bern: Francke, Grammatica storica dell'italiano e dei suoi dialetti (edizione italiana), Fonetica, Torino: Einaudi, 1966, 439-446.

Roland, V., Huet, K., Harmegnies, B., Piccaluga, M., Verhaegen, C. & Delvaux, V. (2023). Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease. In *Frontiers in Psychology*, 14.

Romano, A., Manco, F. (2004). Incidenza di fenomeni di riduzione vocalica nel parlato spontaneo a Bari e a Lecce. In Albano-Leoni, F., Cutugno, F., Pettorino, M. & Savy,

R. (a cura di), *Atti del Convegno Naz. Napoli*, "Il Parlato Italiano", 13-15 Febbraio 2003), Napoli: D'Auria (CD-ROM).

Roy, N., Nissen, S.L., Dromey, C. & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: evidence for vowel space expansion following manual circumlaryngeal therapy. In *Journal of Communication Disorders*, 42, 124-135.

Sapir, S., Ramig, L.O., Spielman, J.L. & Fox, C. (2010). Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech. Journal of Speech. In *Language, and Hearing Research*, 53, 114–25.

Schiel, F. (1999), Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS*, San Francisco, August 1999, 607-610.

Skodda, S., Visser, W. & Schlegel, U. (2011). Vowel Articulation in Parkinson's Disease. In *Journal of Voice*, 25, 467–72.

Skodda, S., Gronheit, W. & Schlegel, U. (2012). Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease. In *PLoS ONE*, vol.7 (2), 1–8.

Siniscalchi, S.M. (2021). Vector-to-Vector Regression via Distributional Loss for Speech Enhancement. In *IEEE Signal Processing Letters*, vol. 28, 254-258.

Tjaden, K. (2000). An acoustic study of coarticulation in dysarthric speakers with Parkinson disease. In *Journal of speech*, language, and hearing research, 43 (2000), 1466-1480.

Vásquez Correa, J.C., Arias, T., Orozco-Arroyave, J.R. & Nöth, E. (2018). A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease. In *Proceedings of Interspeech* 2018, 456-460, doi: 10.21437/Interspeech.2018-1988

Wang, Q., Fu, Y., Shao, B., Chang, L., Ren, K., Chen, Z. & Ling, Y. (2022). Early detection of Parkinson's disease from multiple signal speech: Based on Mandarin language dataset. In *Front Aging Neurosci*. 14: 1036588.

Wong, M.N., Murdoch, B.E., & Whelan, B.M. (2010). Kinematic analysis of lingual function in dysarthric speakers with Parkinson's disease: An electromagnetic articulograph study. In *Int. JS-LPathology*, vol. 12, 414–425.

Wong, M.N., Murdoch, B.E. & Whelan, B.M. (2011). Lingual Kinematics in Dysarthric and Nondysarthric Speakers with Parkinson's Disease. In *Parkinson's Disease*, 1-8.

DOI: 10.17469/O2111AISV000007

CLAUDIO FERRANTE, VINCENZO SCOTTI

Cross-Lingual Transferability of Voice Analysis Models: a Parkinson's Disease Case Study

Traditionally, speech analysis has always relied on a set of very informative features like (Mel) spectrogram, Mel Frequency Cepstral Coefficients (MFCC), pitch or intensity to build speech powered applications. Recently, deep learning-based models for the extraction of acoustic features have allowed significantly improving the state of the art in many speech-related applications. With this work, we focus the analysis on the cross-lingual transferability of speech analysis features. The idea is to understand whether and how well a classification model trained on speech features in a source language works on an unseen target language. We evaluate these properties analysing models for Parkinson's disease detection from speech, adapting the models from English to Telugu. Results show that multi-lingual pre-trained deep learning-based features do not require explicit adaptation and work well out-of-the-box. Differently, models not adapting out-of-the-box respond well even to unsupervised adaptation on a small data set.

Keywords: Natural Language Processing, Deep Learning, Speech Analysis, Parkinson's Disease, Domain Adaptation, Multilingual

1. *Introduction*

Deep learning has become pervasive in Natural Language Processing (NLP), enabling complex language-based applications. Deep neural network models for language and speech analysis are now capable of complex processing like dialogue generation (Gao, Galley, and Li, 2019) and expressive speech synthesis (Tan, Qin, Soong and Liu, 2021). Now the focus is shifting towards multi-task and multi-language support, enabled by data availability and computing power (Raffel. Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li and Liu, 2020; Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever and Amodei, 2020).

These deep learning-based models are powerful end-to-end tools to solve language-related tasks and can also be used for feature extraction. The hidden transformations learnt by these deep learning models can be reused on other tasks involving input data from a similar domain through the transfer learning process. This process consists in reusing the features extracted by the hidden transformations of a pre-trained deep neural network as input to a new classification neural network (Yosinski, Clune, Bengio and Lipson, 2014).

In this paper, we focus on analysing speech features and their cross-lingual transferability. The objective is to evaluate how well a model trained in one language (hereafter referred to as source language), starting from deep learning-based features, can be reused in another language (hereafter referred to as target language) and whether it is necessary to adopt explicit measures to ensure the adaptation goes fine. The objective is to analyse the generality of these features. We selected Parkinson's disease recognition from speech as the target task, and we used English and Telugu, respectively, as source and target languages.

We divide the rest of this paper into the following sections. In § 2 we present the research work connected to the one we are proposing. In § 3 we introduce the neural network-based classification pipeline we adopted. In § 4 we present the available data, as well as the pre-processing and preparation steps. In § 5 and § 6 we describe the experiments we conducted and the results we obtained. Finally, in § 7 we summarise our work and present hints of possible future works.

2. *Related works*

In the following, we introduce the deep neural network models for acoustic features extraction, the existing machine learning-based solutions for Parkinson's disease detection and the concepts behind domain adaptation.

2.1 Neural networks for speech analysis

Traditionally, speech-based classification and recognition tasks were solved using handcrafted prosodic and acoustic features like (Mel-)spectrogram, Mel Frequency Cepstral Coefficients (MFCC), energy, or pitch (Giannakopoulos 2015; Chernykh, Sterling and Prihodko, 2017). Nowadays, deep learning-based features are slightly substituting these handcrafted ones, thanks to the improved performances they are enabling. Usually, learning good deep learning-based features requires large data sets, so this approach is not directly applicable to any task. However, once trained, the feature extraction modules can be reused effectively also on problems with limited data availability. There exist different pre-trained audio and speech analysis models that can be transferred on other tasks.

SoundNet (Aytar Vondrick and Torralba, 2016) is a 1D convolutional neural network for raw audio analysis. The original neural network was trained to mimic the output of video analysis models for object and scene recognition starting from the audio track. The reference vision model would analyse the images from a video to provide the pseudo-labels for classification, and SoundNet would process the audio learning to predict the pseudo-labels from it. The hidden features have been adapted successfully to other classification tasks, like highlight detection in sports events (Merler, Joshi, Mac, Nguyen, Hammer, Kent, Xiong, Do, Smith, and Schmidt Feris, 2018).

VGGish (Hershey, Chaudhuri, Ellis, Gemmeke, Jansen, Moore, Plakal, Platt, Saurous, Seybold, Slaney, Weiss, and Wilson, 2017) is a 2D convolutional neural

network obtained by refining the VGG neural network for image classification. Differently from SoundNet, this model takes as input the audio spectrograms, interpreting them as greyscale images, instead of raw audio waveforms. The model, in this case, was pre-trained on a large audio classification task. As for SoundNet, the hidden features of VGGish have been adapted to other classification tasks like emotion recognition from speech (Scotti, Galati, Sbattella, and Tedesco, 2020).

Wav2Vec and Wav2Vec 2.0 (Schneider, Baevski, Collobert, and Auli, 2019; Baevski, Zhou, Mohamed, and Auli, 2020) differ from the previous two models in training approach: they adopt self-supervised generative training rather than supervised training. This training approach does not require any labelling, allowing to scale the data set size significantly compared to the other approaches. The models are trained to reconstruct the input speech signal from a compressed representation. The second version adopts a state-of-the-art transformer architecture, and both models were developed initially to be refined for Automatic Speech Recognition (ASR).

2.2 Machine learning for Parkinson's disease detection

The approach to Parkison's disease recognition from speech has evolved following the advances of NLP. Initial recognition pipelines used handcrafted features as input and different machine leaning-based algorithms to solve the task, while currently, the approach is to transfer deep learning-based features.

Initial solutions for Parkison's disease detection using speech analysis and machine explored a wide variety of feature combinations and different classification algorithms (Williamson, Quatieri, Helfer, Perricone, Ghosh, Ciccarelli, and Mehta, 2015; Dimauro, Caivano, Bevilacqua, Girardi, and Napoletano, 2016; Dimauro, Di Nicola, Bevilacqua, Caivano, and Girardi 2017; Karan, Sahu, and Mahto, 2020; Rahman, Lee, Islam, Antony, Ratnu, Ali, Al Mamun, Wagner, Jensen-Roberts, Waddell, Myers, Pawlik, Soto, Coffey, Sarkar, Schneider, Tarolli, Lizarraga, Adams, Little, Dorsey and Hoque, 2021). The adopted input features ranged from Mel-spectrogram and MFCC to F0, intensity, jitter, and shimmer. The classification algorithms of choice were mainly Support Vector Machines (Cortes and Vapnik, 1995) and Neural Networks (Goodfellow, Bengio, and Courville, 2016). All approaches applied a similar pre-processing: (i) extracting the feature sequences from the input speech signals, (ii) applying a pooling transformation to remove the temporal dimension from the input and obtaining a feature vector, (iii) optionally applying a dimensionality reduction transformation (e.g., Principal Component Analysis – PCA) to obtain a more compact input vector (Pearson, 1901).

Recently deep learning-based features have also been adopted for Parkison's disease detection, reporting improved performances (more than 80% recognition accuracy) with respect to solutions using standard features (Kurada and Kurada, 2020). Usually, deep learning-based features help yield better results, especially in the presence of large data sets. However, as we also show in this research, traditional prosodic and acoustic features can reach competitive performances. Moreover, traditional and handcrafted features help maintain properties like the explainability

and the interpretability of the results of the classifiers, also on problems like Parkison's disease or Alzahimer's detection (Favaro, Motley, Samus, Butala, Dehak, Oh, and Moro-Velazquez, 2022a; Favaro, Motley, Samus, Butala, Dehak, Oh, and Moro-Velazquez, 2022b).

All the existing works focused on languages individually. Most of the available results come from a benchmark data set for English (more details on this data in § 4). Lately, the attention is shifting towards multilingual solutions. However, even in cases where the classifiers were evaluated on multiple languages, none of the models trained on one language was evaluated on another to assess the cross-language performances (Toye and Kompalli, 2021). The standard approach is to train an ad-hoc model per language. Despite this being the best possible approach, it is not applicable in conditions of scarce data availability.

2.3 Domain adaptation

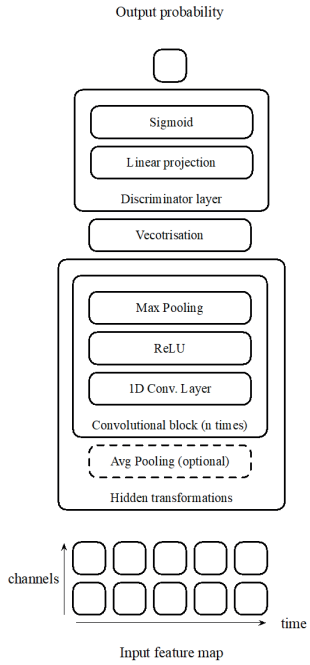
Domain adaptation is the area of machine learning that deals with the problem of applying transfer learning in cases where the original training data (the source data) are slightly different from the new inference data (the target data) (Redko, Morvant, Habrard, Sebban, and Bennani, 2019). We talk of data domain (source and target domains) to indicate the distributions from which samples come from, and of domain shift to address the change in distribution between the source and target domain. We distinguish between unsupervised and supervised domain adaptation depending whether the labels on the target data are unknown or known, respectively. In this work, we focus on unsupervised approaches even though we have target data labels to test our approach in the most complex scenario.

Unsupervised methods are designed to search for a common representation space where samples of the two domains are indistinguishable, while maintaining good performances on the original task. Through the years, many different solutions have been proposed. CORAL (Sun, Feng, and Saenko, 2016) and its deep learning variant Deep-CORAL (Sun and Saenko, 2016), focus on reducing the difference in covariance between the two domains so that the variance of the source and target distributions are indistinguishable. CORAL replaces the covariance of the source data with that of the target prior training, while Deep-CORAL adds an explicit term to the training loss of the neural network it is applied to, so that the difference in the covariance between source and target data is minimised. Gradient reversal layer (Ganin and Lempitsky, 2015) instead uses an adversarial approach and builds a classifier to distinguish between the two domains, and then inverts the gradient updates of the classifier to make the two domains undistinguishable.

Supervised (and sometimes semi-supervised) methods are easier to deal with (Redko et al., 2019). If sufficient target data points are available, it is possible to train directly on that. Otherwise, all unsupervised approaches can be adopted adding the available labels on the target data and computing the regular task loss on those labels alongside the adaptation loss. Alternatively, simple techniques like weighted losses or gradient boosting are sufficient in these settings.

3. *Neural network model*

Figure 1 -*1D convolutional neural network architecture*



In this section we describe the neural network model we used to classify audio samples into. We provide the details of the feature extraction modules to convert the raw audio into a more suitable format, the architecture of the neural network classifier and the training approach.

3.1 Input features

The proposed classifier works on 1D features maps, that are matrices of features $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ where each column-vector $\boldsymbol{x} \in \mathbb{R}^{d}$ corresponds to a specific time stamp, and each value in the vector corresponds to the value of a feature for that time stamp. Slices of features through time (rows of the input matrix) are referred to as channels. The raw input waveform is converted to this feature map, either stacking or using deep learning-based feature extraction pipelines. The model we propose is agnostic of the input features and can be trained with any of the inputs we propose hereafter.

Concerning traditional features, we followed the work of Toye and Kompalli (2021) and adopted the following prosodic and acoustic ones:

- 13 MFCC;
- F0 (pitch);
- Jitter (absolute, relative, rap, and PPQ5);
- Shimmer (absolute, absolute dB, relative, APQ3, and APQ5);
- Harmonicity.

To extract the prosodic and acoustic features, we re-sampled the audio clips at 16kHz and we computed the features in sliding windows with a width of 1024 samples and a hop of 256 samples.

For deep learning-based features, instead, we considered the three main available models for audio and speech features extraction: VGGish, SoundNet and Wav2Vec 2.0. We used all pre-trained implementations of the three neural networks. For the details about the temporal windows, we invite the readers to refer to the original papers.

3.2 Neural network classifier

The classification model $f_{PD}(\cdot)$ we are proposing is a 1D Convolutional Neural Network (CNN). The CNN, which we represented in Fig. 1, is composed by a stack of convolutional blocks, a pooling layer, and a final classification layer. Prior to the first convolutional block, we considered a layer of average pooling to reduce the temporal dimension of the input features. Note that being the CNN a fully convolutional model, it can process inputs of varying and undefined length. Hereafter, we report the general description of the architecture presented in Fig. 1; for further details on the hyperparameters of each block, refer to § 5.

The stack of convolutional blocks is used to learn a hierarchy of hidden features on top of the input ones to solve the classification problem (Parkison's disease detection). Each block is composed of three pieces: a 1D convolutional layer, a $ReLU(\cdot)$ activation and a max pooling transformation to reduce the spatial dimensionality of the input feature map. The output of one block is forwarded to the following one as input. The overall stack of transformations constitutes the hidden transformation $\boldsymbol{h}(\cdot)$ of our CNN, this function maps the input feature map $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ to another feature map $\boldsymbol{H} = \boldsymbol{h}(\boldsymbol{X}) \in \mathbb{R}^{t' \times h}$.

The pooling layer after the convolutional blocks completes the hidden transformation, deleting the temporal dimension of the feature map resulting from the hidden transformations. This step allows to compress all the information in a single vector $\boldsymbol{h} = \boldsymbol{f_{pooling}}(\boldsymbol{H}) \in \mathbb{R}^{h'}$ that can be finally used for the classification. We considered three different pooling transformations: Global Average Pooling (GAP), Global Max Pooling (GMP) and flattening (Lin, Chen and Yan, 2014). With the first two approaches we have $\boldsymbol{h'} = \boldsymbol{h}$, while with the latter $\boldsymbol{h'} = \boldsymbol{t'} \cdot \boldsymbol{h}$.

The final classification layer is composed by a learnable linear projection that maps the hidden vector $\boldsymbol{h}$ to a scalar value, the logit. A final sigmoid activation function $\sigma(\cdot)$ allows to output the probability of the input sample speech being from a subject affected by Parkinson's disease.

3.3 Training approach

The loss function $\mathcal{L}(\cdot)$ optimised to train the neural network is a linear combination of two losses: the former is the usual binary cross-entropy (BCE) loss $\mathcal{L}_{BCE}(\cdot)$, and the latter is the Deep-CORAL objective $\mathcal{L}_{DC}(\cdot)$ to achieve unsupervised domain adaptation. Given the labelled source data $D_s = \{(\boldsymbol{X_{S,1}}, y_{S,1}), \ldots, (\boldsymbol{X_{S,n}}, y_{S,n})\}$,

composed of input-target output pairs (with $y_{S,i} \in \{\mathbf{0}, \mathbf{1}\}\ \forall \boldsymbol{i}$) and the unlabelled target data $D_T = \{\boldsymbol{X_{T,1}}, \ldots, \boldsymbol{X_{T,m}}\}$

$$\mathcal{L}(D_s, D_T) = \mathbb{E}_{(\boldsymbol{X_{S,I}}, y_{S,i}) \in D_s}[\mathcal{L}_{BCE}(\boldsymbol{X_{S,I}}, y_{S,i})] + \lambda \mathcal{L}_{DC}(D_s, D_T)$$

where $\lambda \in [0, 1] \subseteq \mathbb{R}$ is a parameter to control the contribution of the adaptation loss (note that $\lambda = 0$ disables the adaptation).

The binary cross-entropy loss $\mathcal{L}_{BCE}(\cdot)$ is computed as

$$\mathcal{L}_{BCE}(\boldsymbol{X_{S,i}}, y_{S,i}) = y_{S,i} \cdot \log(\boldsymbol{f_{PD}}(\boldsymbol{X_{S,i}})) + (\mathbf{1} - y_{S,i}) \cdot \log(\mathbf{1} - \boldsymbol{f_{PD}}(\boldsymbol{X_{S,i}}))$$

The Deep-CORAL loss for adaptation is computed as

$$\mathcal{L}_{DC}(D_s, D_T) = \frac{1}{4h'^2} \| \Sigma_S - \Sigma_T \|_F^2$$

where $\Sigma_S \in \mathbb{R}^{h' \times h'}$ and $\Sigma_T \in \mathbb{R}^{h' \times h'}$ are the covariance matrices of the hidden feature vectors computed from the samples in the source and target data sets respectively, and $\|\cdot\|_F^2$ is the squared Frobenius matrix norm operator.

4. *Data*

In this section, we present the audio samples we used in this research and all the pre-processing and preparation steps we applied to the available samples.

4.1 Data set

For the scope of this work, we considered two distinct data sets composed by speech samples coming from persons either affected by Parkinson's disease or in healthy conditions. The former data set contains speech recordings on English speakers, while the latter contains speech recordings of Telugu speakers.

The English data set is a known data set for Parkinson's disease detection: the Mobile Device Voice Recordings at King's College London (MDVR-KCL) (Jaeger, Trivedi, and Stadtschnitzer, 2019). The recordings are phone calls sampled at 44.1 kHz 16 bit. Samples come from patients affected by Parkison's disease at different stages of the disease and from patients in healthy conditions. We have samples of female speakers and male speakers from different age ranges (age distribution of patients affected by Parkinson's disease is higher to the disease affecting mostly elderly people). Of the available recordings, 35 m 38 s (corresponding to 31 samples) come from patients affected by Parkinson's disease and 46 m 5 s (corresponding to 42 samples come from patients in healthy conditions.

The Telugu data set is the motivation behind this work. We were provided with a private collection of 12 m 39 s of audio recordings from native Telugu speakers affected by Parkinson's diseases, for a total of 57 samples. Since all the samples were coming from unhealthy patients, we sampled randomly 200 sample (15 m 35 s of recordings) of Telugu speakers in healthy conditions from the Telugu split of the Open Speech and Language Resources (OpenSLR) corpus (He, Chu, Kjartansson,

Rivera, Katanova, Gutkin, Demirsahin, Johny, Jansche, Sarin, and Pipatsrisawat, 2020), a data set of audio books for speech recognition, to obtain the healthy patients' samples. We took half of the samples from male speakers and half from female speakers. We do not have access to the gender information of private collection, and we do not have any information on age distribution of both data sets.

4.2 Pre-processing and preparation

Figure 2 - *Audio samples duration distributions divided per language (the green line is the combine distribution of both kind of patients, values refer to duration after sentence boundaries cropping)*

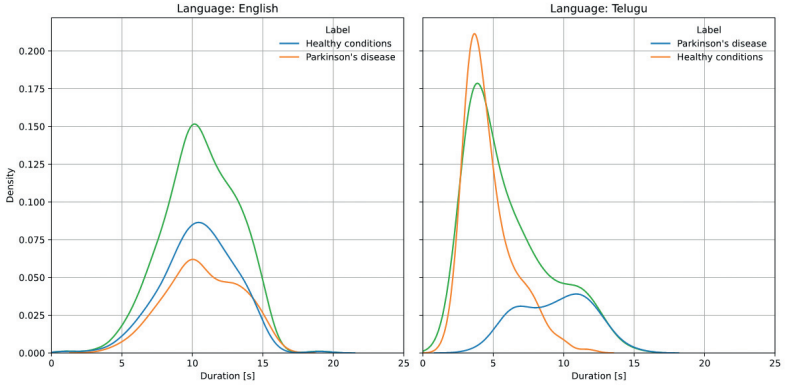


Table 1 - *Audio samples statistics organised per language and per label (PD: Parkinson's Disease, HC: Healthy Conditions) and divided into training and test split, counts are taken after sentence boundaries splitting and after windowing (total counts aggregate across data splits, both counts aggregate across labels)*

Language	*Label*	*Training*		*Test*		*Total*	
		Sentences	*Windowed*	*Sentences*	*Windowed*	*Sentences*	*Windowed*
English	*PD*	148	465	51	149	199	614
	HC	197	591	64	193	261	784
	Both	345	1056	115	352	460	1480
Telugu	*PD*	57	153	24	70	81	223
	HC	153	244	47	71	200	315
	Both	210	397	71	141	281	538

Given the available samples, we applied the same pre-processing steps to both data sets. First, we applied denoising to make sure that only human voice would be present in the recording, then we cropped the audio clips in correspondence of sentence boundaries. Finally, we applied windowing to the resulting samples sample. The neural network classifies individual windows.

To denoise the audio clip we used the RNNoise tool from Mozilla (Valin, 2018). The tool uses a neural network to estimate the spectral components of noise in

each audio recording and applies spectral subtraction to get rid of the noise, while preserving human voice in the recording as clear as possible.

Sentence boundaries splitting was applied manually on both data sets. We looked for long pauses or points where we were sure there was some syntactic boundary, and we divided the original audio sample. We did not apply this process to the OpenSLR part of the Telugu data. The resulting audio samples duration distribution and the actual number of samples resulting from this process are reported, respectively, in Fig. 2 and Tab. 1.

We applied a further windowing to the resulting samples. We cropped the recordings in windows of 4 s with a stride of 2 s between one window and the other. The resulting number of windowed samples is reported in Fig. 2 as well.

5. *Experiments*

We experimented training different configurations of the neural network (changing the input features and the pooling approach) and we compared results on source language as well as results on target language. For the latter we considered the results with and without adaptation. The main experiment was on the adaptation from English to Telugu; however, we also evaluated the other way around for completeness. In both cases, we applied a train-test split on the source language data (75%-25% split) and used the entire target data set for evaluation.

To complete data preparation, we applied standardisation to the input features to have all the feature vectors in the input feature maps have zero mean and unit standard deviation. We estimated separately the statistics for source data (on the training set) and target data (on the entire data set). Additionally, we applied minority oversampling to the source training data to have a balanced training set.

We considered the four different input features introduced in § 3.1 and the three different pooling approaches introduced in § 3.2. For each feature and pooler combination we run a 5-fold cross-validation to find the best configuration. For every fold, we reserved 20% of the training data for validation and early stopping (to prevent overfitting). After finding the best configuration we re-trained the best configuration using all the available source training data (maintain the 20% split for validation).

As training hyperparameters we set the batch size to 16, the maximum number of training epochs to 150 and the early stopping patience to 5. For the neural network we used separable 1D convolution with the kernel size set to 3 and the stride on 1, while for max pooling we set the kernel size to 2 and the stride to 2. Before each convolution, we applied dropout with a rate of 10% for regularisation. We trained all configurations using ADAM as optimiser (Kingma and Ba, 2015). During the cross-validation we tuned the following hyperparameters:

- Number of convolutional blocks: 1, 2, or 3.
- Number of hidden units in the convolutions: 512 or 1024 (we used the same value for all convolutional blocks).
- Learning rate: 10^{-3} or $5 \cdot 10^{-4}$.

Due to the high density in the temporal dimension of traditional features and Wav2Vec 2.0 features, we used the optional initial average pooling layer with a kernel size of 64 elements and a stride of 32.

Concerning adaptation, we used the best configuration of each input feature-pooling approach pair found on the source data also for the adaptation experiment. During training with adaptation, we set the weight of the Deep-CORAL loss to $\lambda = 1/10$.

For evaluation, we computed the usual classification metrics: accuracy, precision, recall, F1, specificity (negative class recall), and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. For the target language, we compared the results with and without the adaptation, to better understand its effects.

6. *Results*

Figure 3 - *English Parkinson's disease detection classifiers results*

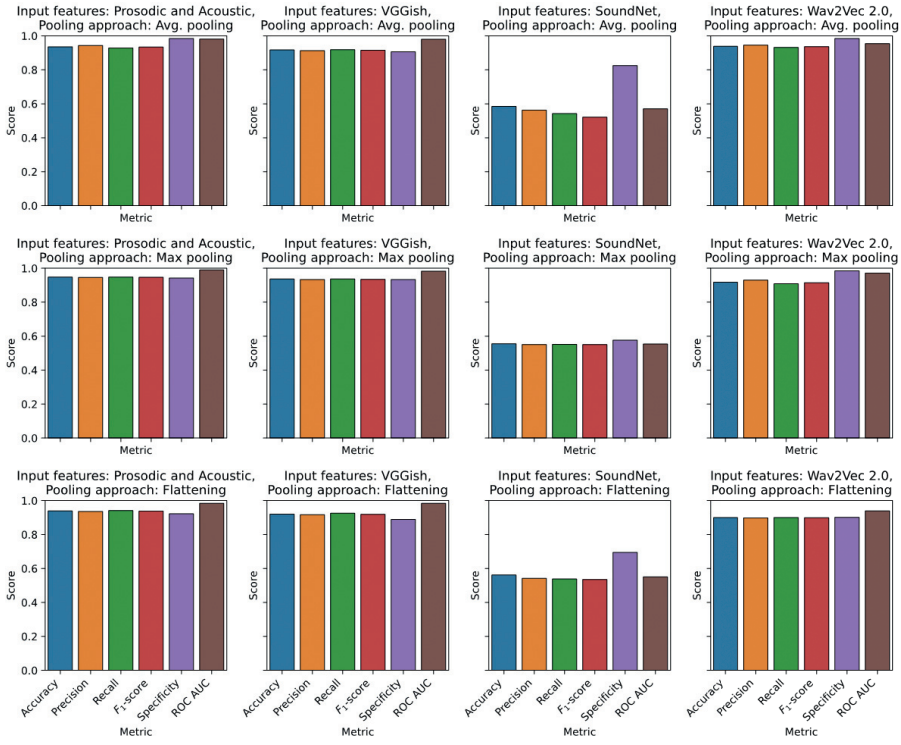
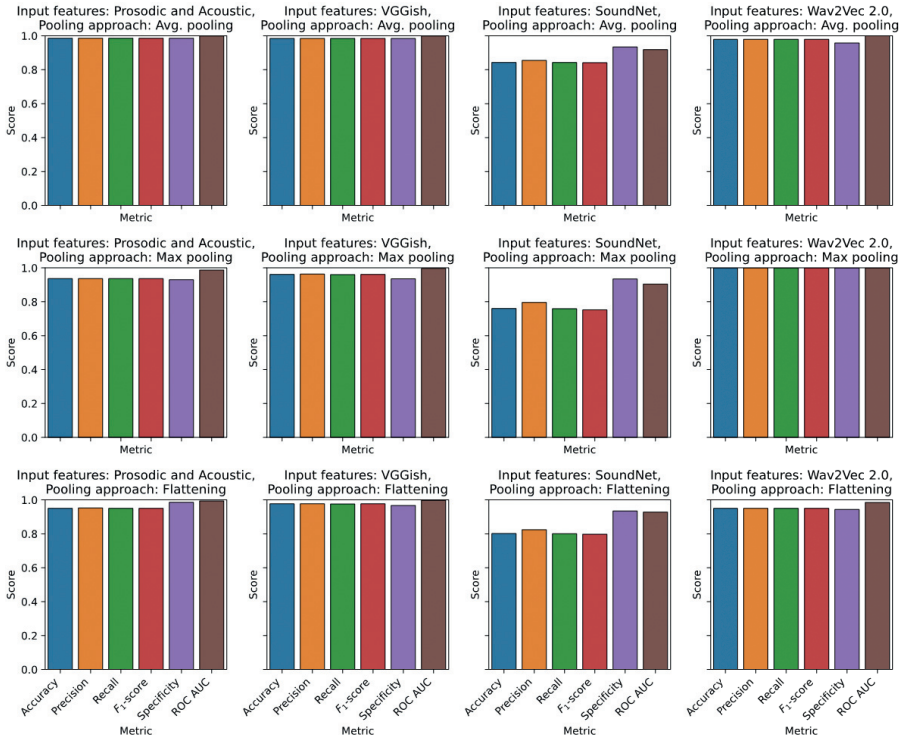


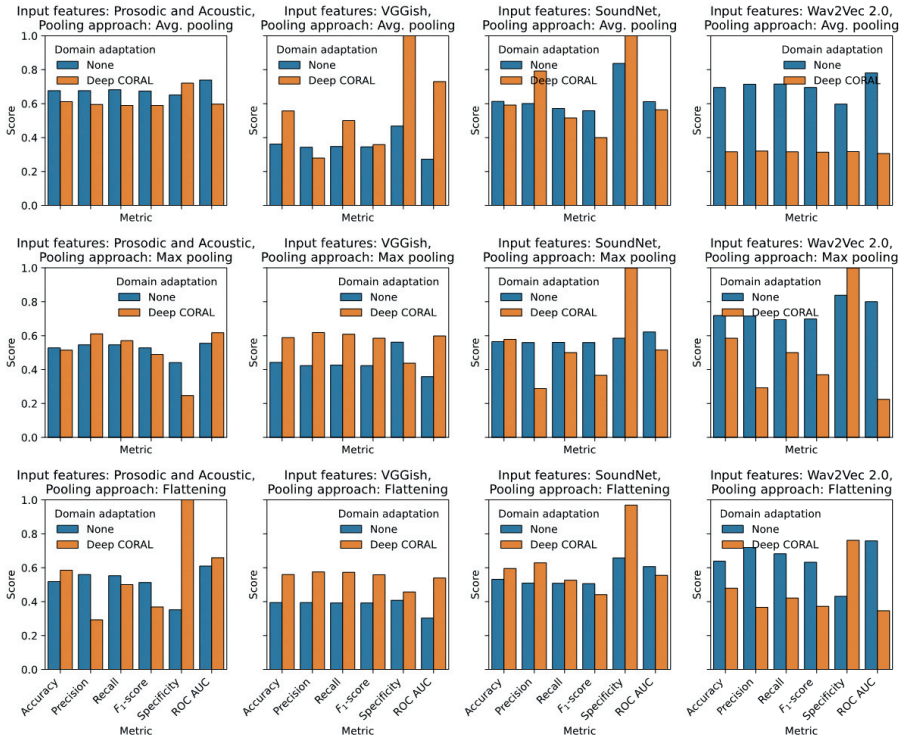
Figure 4 - *Telugu Parkinson's disease detection classifiers results*



We reported the results on the test split of the source languages in Fig. 3 (for English) and Fig. 4 (for Telugu). These results were obtained separately after training on the test split for that same source language.

Results are similar between English and Telugu, although having Telugu fewer samples, they are less robust. On both languages we managed to obtain impressive recognition results. Scores are >90% on all metrics with all input features apart from SoundNet. Traditional features achieve competitive results at the same level of VGGish and Wav2Vec 2.0. About SoundNet, we hypothesised that the lower scores are due to the feature extractor not having "seen" human voice in the pre-training.

Figure 5 - *Comparison of results of domain adaption using Deep-CORAL, on Parkinson's disease detection from English (source language) to Telugu (target language), with respect to no-adaptation baseline*



Concerning adaptation from English to Telugu, we reported the results with and without Deep-CORAL loss in Fig. 5. In general, we obtained mixed results. On traditional features the results with and without adaptation are mostly undistinguishable. Same applies for SoundNet; however, from the source data results analysis, SoundNet features seemed to be useful to this classification task. On VGGish we noticed the expected behaviour, results without adaption were mostly below 50% and Deep-CORAL helped improve the performances on the target language. The result is not as sharp as we expected, but given the unsupervised approach and the reduced data set, it is still valuable. Finally, from Wav2Vec 2.0 we notice the opposite behaviour. The model performed well out-of-the-box with the target language, and the adaptation decreased the performances. We hypothesise that the cross-language results without adaptation are to the self-supervised pre-training of Wav2Vec 2.0 feature extractor, that was conducted on a large language pool (Baevski et al., 2020).

Adaptation from Telugu to English did not highlight any significant result. However, given the reduced size of the source data set, we were not expecting any result in this sense and we run the experiments only for completeness.

7. *Conclusions and future works*

In this paper we explored domain adaptation applications to Parkinson's disease detection, adapting from a source language to a target one. The scope of the research was to understand if, in presence of reduced and unlabelled data sets it would be possible to resort to this kind of techniques. We selected English as source data and Telugu as target data, and we explored different input features to feed a CNN, both traditional prosodic and acoustic features and newer deep learning-based features.

From the results we collected, we noticed that some deep learning-based features obtained via self-supervised pre-training on multiple-languages (Wav2Vec 2.0) are naturally compatible with languages unseen during the transfer learning and are actually badly affected from domain adaptation (results were better when applying an English classifier directly to Telugu rather than using adaptation). However, in the cases where the performances without adaptation were low on the target data, the selected domain adaptation technique correctly helped improve the results (e.g., the case of VGGish based classifiers). Given that we were approaching the problem in a completely unsupervised way, we expect that having access to a larger target data set (even if unlabelled) would help improve the results.

The next steps on this work are to scale up the experiments. We are interested in considering more languages and more adaptation techniques. In fact, we are willing to see whether we can directly train a multilingual model. Moreover, we want to understand if other adaptation techniques can help yield better results.

Acknowledgements

The authors wish to thank Prof. Anitha S. Pillai of Hindustan Institute of Technology and Science for providing the audio samples of Parkinson's disease patients in Telugu.

References

Aytar, Y., Vondrick, C. & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon & Roman Garnett (a cura di), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, Barcelona, Spain, 5-10 December 2016, 892–900. URL https://proceedings.neurips.cc/paper/2016/hash/7dcd340d84f762eba80aa538b0c527f7-Abstract.html

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan & Hsuan-Tien Lin (a cura di), *Advances in Neural Information Processing Systems: 354th Annual Conference on Neural Information Processing Systems* (NeurIPS), December 6-12 2020, virtual, vol. 33, 12449-12460 URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020*). Language models are few-shot learners.* In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (a cura di), *Advances in Neural Information Processing Systems: 354th Annual Conference on Neural Information Processing Systems* (NeurIPS), December 6-12 2020, virtual, vol. 33, 1877-1901, URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac 142f64a-Abstract.html

Chernykh, V., Sterling, G. & Prihodko, P. (2017). *Emotion recognition from speech with recurrent neural networks*. In ArXiv abs/1701.08071. URL http://arxiv.org/abs/1701.08071

Cortes, C. & Vapnik, V. (1995). Support-vector networks. In *Mach. Learn*., 20(3):273–297, doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018

Dimauro, G., Caivano, D., Bevilacqua, V., Girardi, F. & Napoletano, V. (2016). Voxtester, software for digital evaluation of speech changes in parkinson disease. In *IEEE International Symposium on Medical Measurements and Applications, MeMeA,* Benevento, Italy, 15-18 May 2016, 1–6, doi: 10.1109/MeMeA.2016.7533761. URL https://doi.org/10.1109/MeMeA.2016.7533761

Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D. & Girardi, F. (2017). Assessment of speech intelligibility in parkinson's disease using a speech-to-text system In *IEEE Access*, vol. 5, 22199–22208, doi: 10.1109/ACCESS.2017.2762475. URL https://doi.org/10.1109/ACCESS.2017.2762475

Favaro, A., Motley, S., Samus, Q.M., Butala, A., Dehak, N., Oh, E.S. & Moro-Velazquez, L. (2022). A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders. In IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar,532-539, doi: 10.1109/SLT54892.2023.10022435

Favaro, A., Motley, S., Samus, Q.M., Butala, A., Dehak, N., Oh, E.S. & Moro-Velazquez, L. (2022). Artificial intelligence tools to evaluate language and speech patterns in alzheimer's disease. In *Alzheimer's & Dementia*, 18(S2):e064913, doi: https://doi.org/10.1002/alz.064913. URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.064913

Giannakopoulos, T. (2015). Pyaudioanalysis: An open-source python library for audio signal analysis. In *PLOS ONE*, 10(12):1–17, doi: 10.1371/journal.pone.0144610. URL https://doi.org/10.1371/journal.pone.0144610

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Francis R. Bach & David M. Blei, (a cura di), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, Lille, France, 6-11 July 2015, vol. 37, 1180–1189. URL http://proceedings.mlr.press/v37/ganin15.html

Gao, J., Galley, M. & Li, L. (2018). Neural approaches to conversational AI. In Artzi, Y. & Eisenstein, J. (a cura di), *Proceedings of ACL 2018,* Melbourne, Australia, 15-2 July 2018, Tutorial Abstracts, 2–7. doi: 10.18653/v1/P18-5002. URL https://aclanthology.org/P18-5002/

Goodfellow, I.J., Bengio, Y. & Courville, A.C. (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press, ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/

He, F., Cathy Chu, S., Kjartansson, O., Rivera, C., Katanova, A., Gutkin, A., Demirsahin, I., Johny, C., Jansche, M., Sarin, S. & Pipatsrisawat. K. (2020). Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC), Marseille, France, 20-25 June, 6494–6503. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.800

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Channing Moore, R., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J. & Wilson, K.W. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2017), New Orleans, LA, USA, 5-9 March 2017, 131–135, doi: 10.1109/ICASSP.2017.7952132. URL https://doi.org/10.1109/ICASSP.2017.7952132

Jaeger, H., Trivedi, D. & Stadtschnitzer, M. (2019). *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls*, Zenodo, May 2019, URL https://doi.org/10.5281/zenodo.2867216

Karan, B., Sekhar Sahu, S. & Mahto, K. (2020). Parkinson disease prediction using intrinsic mode function based features from speech signal. In *Biocybernetics and Biomedical Engineering*, 40(1):249–264.

Kingma D.P. & Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (a cura di). In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 7-9 May 2015, URL http://arxiv.org/abs/1412.6980

Kurada, S. & Kurada. A. (2020). Vggish embeddings based audio classifiers to improve parkinson's disease diagnosis. In *5th IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (*CHASE 2020), Crystal City, VA, USA, 16-18 December, 9–11, doi: 10.1145/3384420.3431775. URL https://doi.org/10.1145/3384420.3431775

Lin, M., Chen, Q. & Yan, S. (2014). *Network in network.* In Bengio, Y. & LeCun, Y. (a cura di), *2nd International Conference on Learning Representations* (ICLR 2014), Banff, AB, Canada, 14-16 April, 2014. URL http://arxiv.org/abs/1312.4400

Merler, M., Joshi, D., Mac, K.C., Nguyen, Q., Hammer, S., Kent, J., Xiong, J., Do, M.N., Smith, J.R. & Schmidt Feris, R. (2018). The excitement of sports: Automatic highlights using audio/visual cues. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops),* Salt Lake City, UT, USA, 18-22 June 2018, 2520–2523. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w49/html/Merler_The_Excitement_of_CVPR_2018_paper.html

Pearson, K. (1901). Liii. On lines and planes of closest fit to systems of points in space. In *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P.J. (2019) Exploring the limits of transfer learning with a unified text-to-

text transformer. In *Journal of Machine Learning Research*, vol. 21(1:140), 5485-5551. URL http://jmlr.org/papers/v21/20-074.html

Rahman, W., Lee, S., Islam, M.S., Antony, V.N., Ratnu, H., Ali, M.R., Mamun, A.A., Wagner, E., Jensen-Roberts, S., Waddell, E., Myers, T., Pawlik, M., Soto, J., Coffey, M., Sarkar, A., Schneider, R., Tarolli, C., Lizarraga, K., Adams, J., Little, M.A., Dorsey, E.R. & Hoque, E. (2021). Detecting parkinson disease using a web-based speech task: Observational study. *Journal of medical Internet research*, vol. 23(10):e26305, 2021.

Redko, I., Morvant, E., Habrard, A., Sebban, M. & Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.. URL https://www.elsevier.com/books/advances-in-domain-adaptation-theory/redko/978-1-78548-236-6

Schneider, S., Baevski, A., Collobert, R. & Auli, M. (2019). *wav2vec: Unsupervised pre-training for speech recognition*. In Gernot Kubin & Zdravko Kacic (a cura di). In *Proceedings of Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, 15-19 September 2019, 3465–3469. doi: 10.21437/Interspeech.2019-1873. URL https://doi.org/10.21437/Interspeech.2019-1873

Scotti, V., Galati, F., Sbattella, L. & Tedesco, R. (2020). Combining deep and unsupervised features for multilingual speech emotion recognition. In Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J. & Vezzani, R. (a cura di). *In Pattern Recognition, ICPR International Workshops and Challenges* – Virtual Event, January 10-15, 2021, Proceedings, Part II, vol.12662 of Lecture Notes in Computer Science, 114–128, Springer. doi: 10.1007/978-3-030-68790-8_10. URL https://doi.org/10.1007/978-3-030-68790-8_10

Sun, B. & Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. In Hua, G. & Jégou, H. (a cura di). In *Proceedings of the Computer Vision – ECCV 2016 Workshops*, Amsterdam, The Netherlands, 8-10 October 8-10 2016, Part III, vol. 9915 of Lecture Notes in Computer Science, 443–450. doi: 10.1007/978-3-319-49409-8_35. URL https://doi.org/10.1007/978-3-319-49409-8_35

Sun, B., Feng, J. & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In Schuurmans, D. & Wellman, M.P. (a cura di). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 12-17 February 2016, Phoenix, Arizona, USA, 2058–2065. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12443

Tan, X., Qin, T., Soong, F.K. & Liu, T. (2021). A survey on neural speech synthesis. In *CoRR*, abs/2106.15561. URL https://arxiv.org/abs/2106.15561.

Toye, A.A. & Kompalli, S. (2021). Comparative study of speech analysis methods to predict Parkinson's disease. In CoRR, abs/2111.10207. URL https://arxiv.org/abs/2111.10207

Valin, J. (2018). A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *20th IEEE International Workshop on Multimedia Signal Processing (MMSP 2018),* Vancouver, BC, Canada, 29-31 August 2018, 1–5. IEEE, 2018. doi: 10.1109/MMSP.2018.8547084. URL https://doi.org/10.1109/MMSP.2018.8547084

Williamson, J.R., Quatieri, T.F., Helfer, B.S., Perricone, J., Ghosh, S.S., Ciccarelli, G.A. & Mehta, D.D. (2015). Segment-dependent dynamics in predicting parkinson's disease. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 6-10 September, 2015, 518–522. URL http://www.isca-speech.org/archive/interspeech_2015/i15_0518.html

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks?. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., D. & Weinberger, K.Q. (a cura di). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, 8-13 December 2014, 3320–3328. URL https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html

Appendix

For transparency and reproducibility, we release the developed source code via GitHub at the following link: https://github.com/vincenzo-scotti/ voice_analysis_parkinson

DOI: 10.17469/O2111AISV000008

MARTA MAFFIA, LOREDANA SCHETTINO, ROSA DE MICCO,
ALESSANDRO TESSITORE

Disfluencies in Parkinson's Disease. A study on Italian early-stage patients[1]

Although Parkinsonian speech is often described as "disfluent", a detailed description of disruption phenomena in Parkinson's Disease (PD) has not been provided so far. The aim of this study is to identify uses and patterns of disfluencies in early-stage Italian PD subjects. The monological speech of 10 PD patients and 10 age- and sex-matched healthy subjects, all Italian native speakers, was annotated, distinguishing between Forward-Looking Disfluencies (silent pauses, lexical and non-verbal fillers, and prolongations, useful to plan message delivery) and Backward-Looking Disfluencies (repetitions, insertions, deletions, and substitutions, used by the speaker to edit something already uttered). PD and healthy productions were compared on four parameters: the number and frequency of disfluencies; their main functions; the syntactic positioning of the items; the duration of silent pauses, filled pauses, and prolongations. Results highlight the presence of specific uses, types, and characteristics of disfluency phenomena in Italian Parkinsonian speech, even at a very early stage.

Keywords: disfluencies, Parkinson's Disease, repairs, hesitations.

1. *Introduction*

Parkinson's Disease (PD) is the second most common movement disorder, originating in the basal ganglia and resulting from degenerative loss of dopaminergic neurons in the *substantia nigra pars compacta*. It affects more than 2-3 % of the population of the world over 65 years of age (De Lau, Breteler, 2006). Caused by poor muscular activation and coordination, hypokinetic dysarthria is one of PD early symptoms and consists of "abnormalities in the strength, speed, range, steadiness, tone or accuracy of movements required for breathing, phonatory, resonatory, articulatory or prosodic aspects of speech production" (Duffy, 2013: 4).

PD-related dysarthria includes, therefore, a range of voice modifications (increased voice nasality, increased acoustic noise, vocal tremor, harsh and breathy voice quality) and speech impairments, both at the segmental and suprasegmental level: imprecise consonantal articulation and reduced vowel space area, reduced voice intensity, significantly narrower tonal range (mono-pitch), mono-loudness,

[1] This article is the result of the collaboration among the authors. However, for academic purposes, Marta Maffia is responsible for sections 1, 2.1 and 2.2, Loredana Schettino for sections 2.3 and 3. Both are responsible for section 4. Marta Maffia, Rosa De Micco and Alessandro Tessitore are responsible for Data Collection.

impaired speech rate and rhythm, longer silent pauses (see, among others, Darley, Aronson & Brown, 1969; Goberman, Coelho, 2005; Skodda, Visser & Schlegel, 2011; Fivela, Iraci, Sallustio, Grimaldi, Zmarich & Patrocinio, 2014).

Even though Parkinsonian dysarthric speech is usually defined as "disfluent", especially in its advanced stages, the literature still does not provide a clear understanding of the specific characteristics of disrupted PD speech (Goberman, Blomgren, 2003). Recent studies on different languages have focused on the occurrence of stuttering-like disfluencies (i.e., one-syllable word repetitions, sound and syllable repetitions, sound prolongations, and blocks) in mild-to-severe PD patients (Goberman, Blomgren & Metzger, 2010; Juste, Sassi, Costa & de Andrade, 2018). They were examined in different speech tasks and their productions were compared to those of healthy speakers and individuals with developmental stuttering. Results seem to support the relationship between stuttering and the functions of the basal ganglia and dopamine, reporting significantly greater disfluency percentages in PD patients than in the healthy control group, especially in the case of within-words disfluencies (stuttering occurring on part of a word) and in the monologic speech task.

A significative effect of the speech task on the quantity and quality of disfluency phenomena was also reported in a case study on a severely disfluent PD patient, native speaker of American English (Van Lancker Sidtis, Cameron & Sidtis, 2012). Disfluent speech was quantified and compared for conversation, two forms of repetition, reading, recited speech, and singing: while phonetic and syllabic dysfluencies appeared across all tasks, due to an impairment in motor planning of speech segments, lexical dysfluencies, related to linguistic planning, were found to be more frequent in spontaneous speech.

The results of another research, conducted on 10 Polish PD patients, reported no differentiation in the speech disfluency severity between different tasks, thus supporting the essentially organic background of disfluency phenomena in PD speech (Półrola, Góral-Półrola, 2015). However, a positive correlation between the frequency of disfluencies and the duration of the disease was found, as also observed for repetitive speech phenomena (both hyperfluent and dysfluent) in a study conducted on 53 PD patients by Benke, Hohenstein, Poewe & Butterworth (2000).

The effect of Levodopa medication on disfluencies has also been studied, to refute or support the so-called "excess dopamine hypothesis", according to which increased levels of dopamine should lead to the development of stuttering in PD speakers, with controversial results (Im, Adams, Abeyesekera, Pieterman, Gilmore & Jog, 2019).

More generally, early studies on speech errors and disfluencies were driven by the observation that these elements could work as predictors of cognitive impairment in pathologies, such as dementia and Alzheimer's disease (König, Satt, Sorin, Hoory, Toledo-Ronen, Derreumaux, Manera, Verhey, Aalten, Robert & David, 2015; Wang, Lian, Pan, Yan, Zhu, Ng, Wang & Yan, 2019). Being aware that typical speech also commonly includes phenomena like repetitions, pauses,

and self-repairs, the first classifications of the phenomena ascribable to the category of "disfluencies" were defined trying to distinguish between disfluencies in typical and atypical speech (Johnson, 1961; Wingate, 1987). Later on, numerous studies concerning different languages have shown that disfluency phenomena are quite normal and pervasive in human spontaneous speech with frequencies of occurrence approximately averaging between 6 to 10 occurrences per 100 words (see Lickley, 2015). Moreover, it has been acknowledged that speakers may use some of these phenomena to manage and monitor their own speech production. Indeed, speakers may use silences, lengthen segmental material, or produce non-verbal or lexical fillers as a means of gaining extra time when they need it for the planning and construction of the upcoming message. Moreover, speakers may decide to edit or repair part of their already produced message through deletions, substitutions, or insertions (Allwood, Nivre & Ahlsén, 1990; Ginzburg, Fernández & Schlangen, 2014). Hence, some of the elements traditionally included in the heterogeneous class of disfluencies may actually provide valuable information on speech planning and discourse structuring and have been thus described as "speech management phenomena" (Allwood *et al.*, 1990).

Given these observations, investigating differences between the patterns of disfluency phenomena in PD and in typical speech may provide a privileged perspective for distinguishing and analysing uses of disfluencies phenomena progressively acquired with the outbreak of the disease and associated with cognitive, linguistic, and motor deficits resulting from the damages in the central nervous system.

In this framework, the present study aims at providing preliminary observations on the specific characteristics of speech disfluency phenomena produced by Italian subjects with early-stage PD. The starting hypotheses in this regard are essentially two: that monological spoken productions of early-stage PD patients and those of healthy individuals may differ with respect to disfluency phenomena; that this difference does not have to do with frequency but with functional, positional and/or durational aspects of disfluent speech.

2. *Method*

2.1 Participants

The data for the present research were collected from 20 Italian native speakers residing in the Campania region: 10 participants with idiopathic non-demented PD were involved (5 males, 5 females; mean age= 65), along with 10 age-matched Healthy Controls (HC, 6 males, 4 females; mean age= 67).

The patients were recruited at the Movement Disorders Unit of the First Division of Neurology at the University of Campania "Luigi Vanvitelli", from a cohort of subjects with no history of previous language and speech disorder. The diagnosis of PD was based on the modified diagnostic criteria of the UK Parkinson's Disease Society Brain Bank (Gibb, Lees, 1988). Inclusion criteria were as follows:

(1) PD onset after the age of 40 years, to exclude early-onset parkinsonism; (2) a modified Hoehn and Yahr (mH&Y) stage ≤ 2.5; and (3) disease duration ≤ 4 years. Exclusion criteria were as follows: (1) relevant cognitive impairment associated with PD assessed with the Montreal Cognitive Assessment (MoCA, Folstein, Folstein & Mchugh, 1975); (2) major depression, minor depression, and dysthymic disorder according to DSM-V criteria, rated by means of the Beck Depression Inventory (Beck, Ward, Mendelson, Mock & Erbaugh, 1961); and (3) any other neurological disorder or clinically significant medical condition.

All patients underwent an extensive assessment with PD-related scales. The presence and the severity of motor impairments were assessed through the Unified Parkinson's Disease Rating Scale part III (UPDRS III), which includes the evaluation of clinically significant speech difficulties (item III.1). The enrolled patients presented mild to moderate motor (speech) impairments at the moment of data collection.

The demographic and clinical data of patients with PD, together with the characteristics of the HC group, are reported in Table 1.

Table 1 - *Demographic and clinical features of patients with PD and HC*

		PD (n= 10) *Mean ± SD*	*HC (n= 10)* *Mean ± SD*
Demographic data	Age	65±11	67±6
	Sex (M/F)	5/5	6/4
	Disease duration (months)	31.7±18.9	
Clinical data	mH&Y stage	1.9±0.5	
	UPDRS III	25.0±8.2	
	MoCA total	24.4±4.3	
	BDI	5.2±6.2	

H&Y: Hoehn & Yahr; UPDRS: Unified Parkinson's Disease Rating Scale; MoCA: Montreal Cognitive Assessment; BDI: Beck depression inventory

2.2 Speech data collection

All participants were asked to perform a monologic task and they were recorded on a standard personal computer in a quiet room at University "Luigi Vanvitelli," by means of the software Praat (Boersma, Weenink, 2021) at a 44,100 Hz sampling rate. PD patients and HC speakers were asked to talk about positive and negative aspects of the place where they lived at the moment of the recording. All the subjects were invited to speak in their normal, conversational voice and at comfortable loudness. Sociolinguistic information on each speaker was also obtained through a questionnaire and all subjects gave written consent to the data collection procedure.

This study concerns the analysis of a subset of approximately 22 minutes of PD and 17 minutes of healthy speech which is part of a larger corpus. The latter amounts to approximately 105 minutes speech including read and spontaneous

monologic speech and was already the object of a rhythmical analysis in a previous work (Maffia, De Micco, Pettorino, Siciliano, Tessitore & De Meo, 2021), which highlighted a comparable articulation rate in the two groups of speakers. The articulation rate was calculated by measuring the VtoV parameter, i.e., the mean duration of the intervals between two consecutive vowel onsets (Pettorino, Maffia, Pellegrino, Vitale & De Meo, 2013): in the PD speech the mean VtoV value was 0.188 s, in healthy speech it was 0.185 s. Moreover, mean data on speech time composition showed no statistical difference in the percentage of disfluent time on the total utterance (PD= 14 %; HC= 11 %). Starting from these observations, the present work aims at providing a more detailed analysis of the way PD and HC speakers may use disfluency phenomena to manage their speech production.

2.3 Analysis

The objects of the analysis are speech elements that are commonly referred to as "disfluencies", though this is a highly debated umbrella term used for a wide range of phenomena. In this study, we observe elements that have been described as "speech management phenomena" as they may be used by speakers to monitor and effectively manage the online processes of speech planning, coding, articulation, and reception (Levelt, 1989). As more extensively described in (Schettino, Betz, Cutugno & Wagner, 2021; Schettino, 2022), each phenomenon is identified and annotated on three main levels[2], with the context of occurrence being discriminative for the identification and annotation processes[3]. On the first level, the disfluencies' macro-structure is labelled: the region to be repaired (*Reparandum*, RM), the repaired one (*Reparans*, RS), and the one where the delay occurs (*Interregnum*, IM). The second level is for the identification of specific items: *Deletions* (DEL), *Insertions* (INS), *Substitutions* (SUB), *Repetitions* (REP), *Silent Pauses* (SP), *Prolongations* (PRL), *Filled Pauses* (FP), and *Lexicalized Filled Pauses* (LFP). On the third level, each item is assigned its main function: *Forward-Looking* (FLD), when their occurrence suspends the speech delivery thus gaining time for speech planning processes; *Backward-Looking* (BLD), when they trace back to and alter speech strings that have been already uttered.

The analysis concerns the comparison of Parkinsonian (PD) and Health Control (HC) speech regarding the following parameters:

- the frequency of disfluent items;
- the main contextual function of the occurrences (BLD, FLD);
- the syntactic positioning of the items, i.e.,
  - within words (WTH – W), e.g., «ma io no+ <ehm> non ho i cani»,
  - within phrases (INTRA – S), e.g., «la zona dei paesi<ii> scandinavi»,

[2] The data were transcribed following the guidelines provided by Savy (2005).

[3] Accordingly, no fixed minimum duration threshold was set to detect neither silent pauses nor prolongations. Then latter were defined as silences or segmental lengthening perceived as marked in their context of occurrence. The reliability of such system was verified by measuring the agreement between two expert annotators (Cohen's k = 0.82, good agreement, Landis, Koch, 1977; in Schettino, 2022).

- between phrases (INTER – S), e.g., «ci sono <mh> <sp> molti monumenti»,
- between clauses (INTER – C), e.g., «ci stanno le persone <eeh> che ti conoscono»;

– the duration of Silent Pauses, Filled Pauses, and Prolongations.

The annotation was conducted using the ELAN software for multi-level linguistic annotations (Sloetjes, Wittenburg, 2008). Considering the high degree of individual variability that characterizes speech productions, the statistical significance of the results is tested by building different Generalised Linear and Linear Mixed Models (lme4, Bates, Maechler, Bolker & Walker, 2015) to observe the relevant and consistent features net of such variability. The models were built by considering the previously listed analysis parameters (frequency of occurrence, main contextual function, syntactic positioning, duration) as dependent variables. The effect of the health condition (PD or HC), the subjects' biographical data (sex and age), and clinical variables (mH&Y stage, UPDRS III, MoCA total, BDI) on the chosen parameters was tested by including them in the models as "explanatory" independent variables. Then, since the analysis is based on the occurrences of disfluency phenomena and each speaker produced definitely more than one phenomenon, the models included Speaker as a random effect to control for individual variability.

3. *Results*

3.1 General frequencies and main function

The analysis concerned 581 disfluency phenomena. In the observed data, they were slightly more frequent in the PD productions (294) than in the HC productions (281).

Also, when considering the different types of phenomena (Figure 1), the most frequent are repetitions, prolongations, and silent pauses. Comparing PD and HC speakers' productions, PD speakers only seem to produce more substitutions and repetitions, but no significant difference emerges (Est.=0.28, SE=2,71, z=0,11, p=0.91). Nonetheless, when considering the main function associated with each item (Figure 2), both PD and HC speakers produce considerably more Forward-Looking Disfluencies than Backward-Looking ones, but PD speakers are found to produce significantly more Backward-Looking Disfluencies than the speakers of the control group (Est.=0.87, SE=0.41, z=2.1, p=0.035). No significant effect is exerted by the biographical variables (sex and age). As for the clinical variables, among PD, cognitive scores significantly affect the use of BLD, the latter are significantly more with lower scores (MoCA, Est.=0.11, SE=0.04, z=2.74, p= 0.006, see Figure 3).

Figure 1 - *Frequency of occurrence of the different types of phenomena per condition and grouped by main function: Backward-Looking (left panel), Forward-Looking (right panel)*

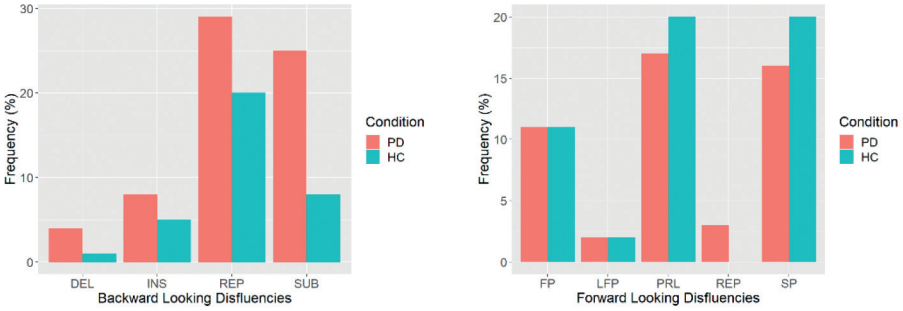


Figure 2 - *Frequency of occurrence of Backward-Looking and Forward-Looking phenomena per condition*

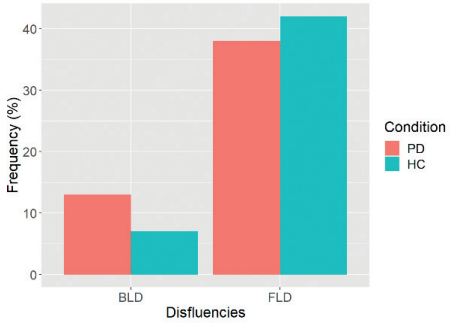
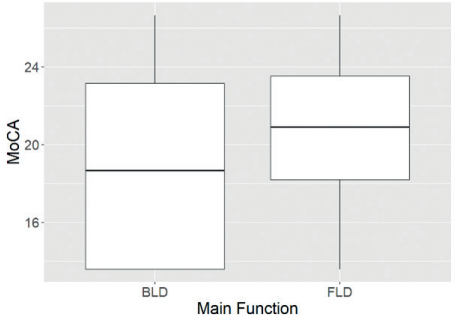


Figure 3 - *MoCA score of Backward-Looking and Forward-Looking phenomena*



3.2 Positioning

The distribution of phenomena with reference to syntactic structures (clauses, phrases, and words) also seems quite different between the PD and the HC groups (Figure 4). In this case, the statistical analysis was conducted by fitting GLMMs with the positioning considered as binomial dependent variable. The model where the outcome had one level for within-word occurrences and the other level grouped the

occurrences in the other positions yielded significant results for both Condition and Main Function. Namely, disfluency phenomena occur significantly more frequently within words, that is interrupting word production when produced by PD speakers (Est.=1,99, SE=0,67, z=2,94, p=0.003) and when associated with a BLD function (Est.=1,6, SE=0,46, z=3,41, p=0.0006). Since the main function associated with disfluency phenomena was found to exert an effect on the positioning of the occurrences, we further tested BLD and FLD separately. Backward-Looking phenomena occur significantly more frequently within words and less frequently between words in PD productions than in HC speech (Est.=1,93, SE=0,71, z=2,73, p=0.006). Similarly, Forward-Looking phenomena interrupt words significantly more frequently in PD than in HC speech (Est.=2,49, SE=0,61, z=4,06, p=<.0001).

Figure 4a-d - *Position of Backward-Looking and Forward-Looking phenomena produced by PD and HC speakers*

Figure a - *PD – BLD*

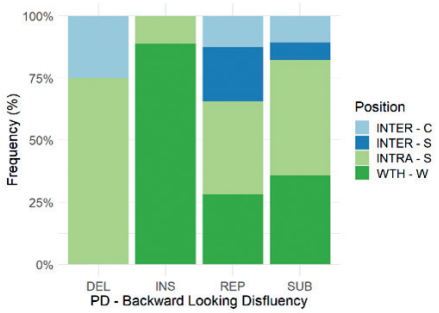


Figure b - *HC – BLD*

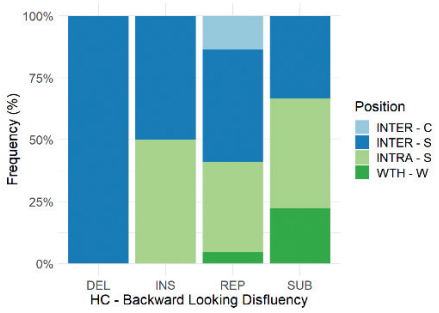


Figure c - *PD – FLD*

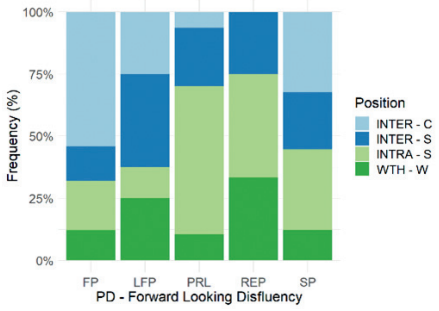
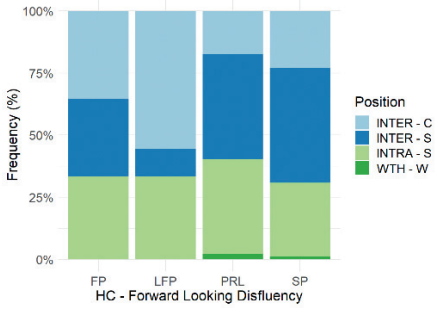
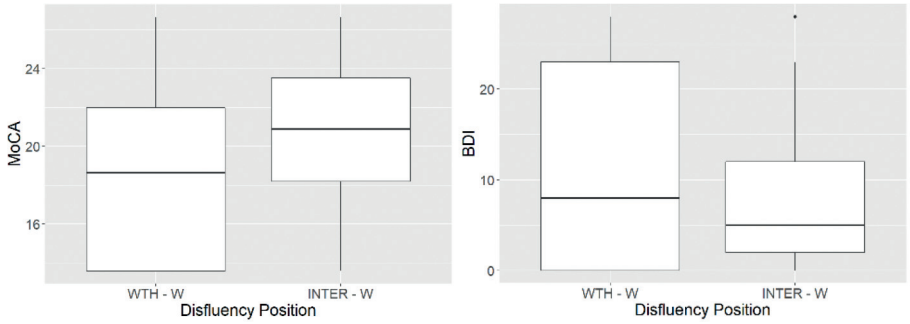


Figure d - *HC – FLD*



Considering the clinical variables among PD (Figure 5), position was found to be affected by cognitive scores (MoCA), phenomena interrupting words being significantly more frequent with lower cognitive scores and phenomena non-interrupting words being more frequent with higher scores (Est.=0.19, SE=0.06, z=3.19, p=0.001). Conversely, within-words phenomena were significantly more frequent with higher depression scores (BDI, Est.=-0.10, SE=0.04, z=-2.3, p=0.02).

Figure 5 - *MoCA score (left) and BDI score (right) per phenomena within-word or between-word position*



3.3 Duration

As reported in Table 1 and illustrated in Figure 6, on average, Forward-Looking phenomena produced by PD speakers tend to be slightly longer than those produced by HC speakers. However, when considering the different types of phenomena, a significant difference only emerged for Silent Pauses (Est=211.28, SE=78.27, t =2.699).

Figure 6 - *Duration of Silent Pauses, Prolongations and Filled Pauses per condition*

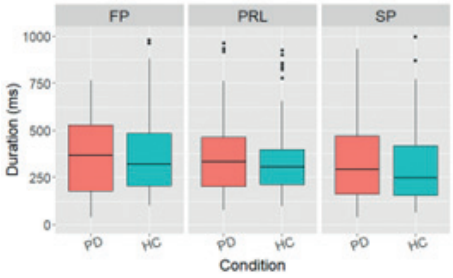


Table 2 - *FLDs duration values per condition: count, mean, standard deviation, standard error of the mean, and confidence interval (default 95 %)*

FLD	*Cond*	*n*	*Dur(ms)*	*sd*	*SE*	*ci*
SP	PD	74	521	555	64	128
SP	HC	89	343	266	28	100
PRL	PD	77	372	372	25	50
PRL	HC	92	344	344	19	38
FP	PD	50	472	360	51	102
FP	HC	50	468	352	49	100

4. *Discussion*

This work has proposed an approach to describe the characteristics of disfluency in Italian patients with early-stage PD by considering the way speakers use disfluency phenomena to manage their spontaneous speech. More specifically, we distinguished between Forward-Looking and Backward-Looking Disfluencies, the former associated with a suspension of message delivery and a need for extra time for language planning, the latter used to edit something that has already been said.

As previously hypothesized, this distinction highlighted significant differences, although the considered PD and HC speech did not differ much in the number of occurrences of speech disfluency phenomena, as also observed by Maffia *et al.* (2021), and both PD and HC speakers were found to more frequently resort to phenomena suspending the speech delivery. In fact, the patients needed to repair something already uttered significantly more frequently than healthy subjects. In particular, they more frequently produced Substitutions, to edit something already uttered and deemed inappropriate to a certain extent, and Repetitions with a bridging function (Shriberg, 1995).

Relevant differences between the production of disfluency phenomena in the PD and HC groups of speakers also concerned the syntactic positioning of disfluencies: while healthy speakers more frequently produce these phenomena between larger units such as clauses and between phrases, PD speakers interrupt words with within-word phenomena significantly more frequently, which corroborates previous findings in the literature (Juste *et al.*, 2018; Goberman *et al.*, 2010).

Moreover, PD patients were observed to produce slightly fewer Silent Pauses and Prolongations, but significantly longer pauses than healthy speakers, which is also in line with previous findings (Darley *et al.*, 1969; Goberman, Coelho, 2005).

At a closer inspection, some clinical variables (cognitive and depression scores mainly) were also found to play a significant role in explaining the observed tendencies in the characterization of PD speakers' production of disfluency phenomena. Although all patients scored above the MoCA threshold and below the BDI threshold that assesses the outbreak of dementia or depression respectively, the increased need for repair observed in PD patients, as compared to healthy speakers, is greater the worse the cognitive condition assessment is. Just as well, PD speakers' need to interrupt words is greater the lower cognitive and the higher depression scores are. These findings support the idea that the features observed as characteristic of PD speech can be considered as a result of a progressive process of deterioration of physiological conditions which determines the development of different strategies to manage the own speech production.

5. *Conclusions*

To conclude, the preliminary findings reported in this study have highlighted that it is the study of the way disfluency phenomena are used by speakers to manage their

speech rather than mere frequency observations that may provide relevant information concerning the features of Parkinsonian speech and the conditions of patients.

It seems that, even at a very early stage of the disease, when patients' speech is still completely intelligible and the effects of dysarthria are not yet perceptible, both motor and linguistic planning and processing are somehow already altered, as the nature and the occurrence of disfluencies reveal. When compared to healthy speakers, PD patients need and spend more time repairing their own utterances, often interrupting the production of words and producing significantly longer silent pauses.

A plausible explanation has to be sought at the intersection of cognitive impairments and motor modifications in speech production, which both seem to affect communication abilities from the very beginning of Parkinson's Disease and even in the absence of dementia (Rohl, Gutierrez, Johari, Greenlee, Tjaden, Roberts, 2022). However, in light of the limited sample size, the results of this study require caution in their interpretation and need further and more detailed replications.

Bibliography

Allwood, J., Nivre, J. & Ahlsén, E. (1990). Speech management – on the nonwritten life of speech. In *Nordic Journal of Linguistics*, 13, 3-48. https://doi.org/10.1017/S0332586500002092

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. In *Arch. Gen. Psychiatry*, 4, 561-571.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Benke, T., Hohenstein, C., Poewe, W. & Butterworth, B. (2000). Repetitive speech phenomena in Parkinson's disease. In *Journal of Neurology, Neurosurgery & Psychiatry*, 69(3), 319-324.

Darley, F.L., Aronson, A.E. & Brown, J.R. (1969). Cluster of deviant speech dimension in the dysarthrias. In *Journal of Speech and Hearing Research*, 12(3), 462-469.

de Lau, L.M. & Breteler, M.M.B. (2006). Epidemiology of Parkinson's disease. In *The Lancet. Neurology*, 5(6), 525-535.

Gili Fivela, B., Iraci, M., Sallustio, V., Grimaldi, M., Zmarich, C. & Patrocinio, D. (2014). Italian vowel and consonant (co) articulation in Parkinson's Disease: Extreme or reduced articulatory variability?. In *Proceedings of 10th ISSP*, May 5-8, Cologne, Germany, 146-149.

Folstein, M.F., Folstein, S.E. & McHugh, P.R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. In *J. Psychiatr. Res*, 12, 189-198.

Gibb, W.R., Lees, A.J. (1988). A comparison of clinical and pathological features of young- and old-onset Parkinson's disease. In *Neurology*, 38, 1402-1406.

Gili Fivela, B., Iraci, M., Sallustio, V., Grimaldi, M., Zmarich, C. & Patrocinio, D. (2014). Italian vowel and consonant (co) articulation in Parkinson's Disease: Extreme or reduced articulatory variability?. In *Proceedings of 10th ISSP*, May 5-8, Cologne, Germany, 146-149.

Ginzburg, J. Fernandez, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. In *Semantics and Pragmatics*, 7(9), 1-64.

Goberman, A.M., Blomgren, M. (2003). Parkinsonian speech disfluencies: effects of L-dopa-related fluctuations. In *Journal of fluency disorders*, 28(1), 55-70.

Goberman, A.M., Coelho, C.A. (2005). Prosodic characteristics of Parkinsonian speech: The effect of levodopa-based medication. In *Journal of Medical Speech-Language Pathology*, 13, 51-68.

Goberman, A.M., Blomgren, M. & Metzger, M. (2010). Characteristics of speech disfluency in Parkinson Disease. In *Journal of Neurolinguistics*, 23(5), 470-478.

Im, H., Adams, S., Abeyesekera, A., Pieterman, M., Gilmore, G. & Jog, M. (2019). Effect of Levodopa on speech dysfluency in Parkinson's disease. In *Movement disorders clinical practice*, 6(2), 150-154.

Johnson, W. (1961). Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. In *The Journal of speech and hearing disorders*, (Suppl.7), 1-20.

Juste, F.S., Sassi, F.C., Costa, J.B. & de Andrade, C.R.F. (2018). Frequency of speech disruptions in Parkinson's Disease and de-velopmental stuttering: A comparison among speech tasks. In *PLoS One*, 13(6), e0199054.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, H.P. & Renaud, D. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's Disease. In *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1), 112-124.

Landis, J., Koch, G. (1977). The measurement of observer agreement for categorical data. In *Biometrics*, 33(1), 159-174. https://doi.org/10.2307/2529310

Levelt, W.J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press. https://doi.org/10.1016/0010-0277(83)90026-4

Lickley, R.J. (2015). Fluency and Disfluency. In Redford, M.A. (Ed.), *The handbook of speech production*. John Wiley & Sons, 445-474. https://doi.org/10.1002/9781118584156.ch20

Maffia, M., De Micco, R., Pettorino, M., Siciliano, M., Tessitore, A. & De Meo, A. (2021). Speech rhythm variation in early-stage Parkinson's disease: a study on different speaking tasks. In *Frontiers in Psychology*, 12, 668291.

Pettorino, M., Maffia, M., Pellegrino, E., Vitale, M. & De Meo, A., (2013). VtoV: A perceptual cue for rhythm identification. In *Proceedings of the Prosody-Discourse Interface Conference*, Leuven, Belgium, September 11-13, 101-106.

Półrola, P.J., Góral-Półrola, J. (2015). Speech disfluencies in Parkinson's disease. In *Medical Studies/Studia Medyczne*, 31(4), 267-270.

Rohl, A., Gutierrez, S., Johari, K., Greenlee, J., Tjaden, K. & Roberts, A., (2022). Speech dysfunction, cognition, and Parkinson's Disease. In *Progress in Brain Research*, 269(1), 153-173.

Savy, R., (2005). Specifiche per la trascrizione ortografica annotata dei testi. In F. Albano Leoni & R. Giordano (Eds.), *Italiano parlato. Analisi di un dialogo (CD-ROM)*. Napoli: Liguori.

Schettino, L. (2022). *The role of disfluencies in Italian discourse. Modelling and speech synthesis applications*. PhD Dissertation, Università degli Studi di Salerno, Italia.

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021). Hesitations and individual variability in Italian tourist guides' speech. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E. & Schmid, S. (Eds.), *XVII Convegno Nazionale AISV. Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*. Studi AISV 8, 243-262.

Shriberg, E.E. (1995). Acoustic properties of disfluent repetitions. In *Proceedings of the international congress of phonetic sciences*, 4, 384-387.

Sloetjes, H., Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *Proceedings of the 6th international Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Marocco, 28-30 May 2008, 816-820.

Skodda, S., Visser, W. & Schlegel, U. (2011). Vowel articulation in Parkinson's disease. In *Journal of voice: official journal of the Voice Foundation*, 25(4), 467-472.

Van Lancker Sidtis, D., Cameron, K. & Sidtis, J.J. (2012). Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech. In *Clinical linguistics & Phonetics*, 26(8), 695-711.

Wang, T., Lian, C., Pan, J., Yan, Q., Zhu, F., Ng, M.L., Wang, L., Yan, N. (2019). Towards the speech features of mild cognitive impairment: Universal evidence from structured and unstructured connected speech of Chinese. In *Proceedings of INTERSPEECH 2019*, 3880–3884.

Wingate, M.E. (1987). Fluency and disfluency; illusion and identification. In *Journal of Fluency Disorders*, 12(2), 79-101.

DOI: 10.17469/O2111AISV000009

VALENTINA COLONNA, FEDERICO LO IACONO, ANTONIO ROMANO

Amelia Rosselli e l'angoscia del respiro: tra stile e riflessi di parlato patologico

Amelia Rosselli and the anguish of breathing: stylistic choices and traces of pathological speech

This research aims to delve into a distinctive and compelling example of poetic discourse, marked by features that may imply an affinity with the domain of pathological speech. For the study were selected eight poetic readings by Amelia Rosselli (1930-1996), one of the most renowned female poets of the Italian literary tradition. The interest in studying the (potential) poet's pathological speech arises first of all from the reconstruction of her clinical history, which finds a reflection in the analysed oral data. The recordings cover a period from 1985 to 1990 and reflect the last phase of Rosselli's brief life. The analysis reveals interesting information on the nature of her speech, offering insights into the possible prosodic reflections of a pathological substrate. Her reading style is characterized by fragmentation through emphasized silence and noise, as well as a peculiar use of pauses. Nevertheless, the stylistic, pathological, or potentially dual nature of Rosselli's discourse remains uncertain. This study employed the *VIP-Voices of Italian Poets* methodology, and a *General VIP-Radar* was constructed to summarize her distinctive style.

Keywords: Amelia Rosselli, pathological speech, Parkinson's disease, phonetics, oral poetry.

1. *Introduzione*

Il presente lavoro ha come oggetto centrale il parlato poetico di Amelia Rosselli (1930-1996), una delle poetesse più importanti e influenti della tradizione poetica italiana novecentesca. Il nostro interesse per questo caso di studio si è sviluppato e ampliato negli ultimi anni, fornendo un primo approccio di indagine in Colonna (2021a; 2022) e poi uno studio approfondito in Lo Iacono (2023). Condotti entrambi applicando la metodologia del progetto *VIP-Voices of Italian Poets*, questi studi hanno fatto emergere dati stimolanti e hanno messo in luce alcuni tratti stilistici caratteristici del modo di leggere rosselliano, tra i quali l'impiego di allungamenti di contoidi e un uso sperimentale della voce in relazione al testo, tale da consentirne l'annoveramento tra le voci "sperimentali" nella storia della lettura del Novecento italiano della poesia (Colonna, 2021a; 2022). Il *modus legendi* rosselliano, che assorbe anche le influenze dello sperimentalismo musicale, si caratterizza, inoltre, per una tendenza alla frammentazione tramite silenzio e rumore enfatizzati dalla microprosodia e da un uso peculiare delle pause. Proprio a partire dall'analisi della pausazione nella lettura rosselliana abbiamo potuto riscontrare come in essa siano

presenti alcune caratteristiche che – per quanto considerate inizialmente soltanto stilistiche – potrebbero, invece, essere il potenziale riflesso di un parlato patologico o, ancor più probabilmente, un'intricata commistione delle due dimensioni. Dunque, alla luce di studi rivolti al silenzio in ambito patologico (cfr. Dovetto, 2013 e Nodari & Calamai, 2021), e considerata inoltre la particolarità stilistica di Rosselli, abbiamo svolto una ricerca sulle pause impiegate dall'autrice nelle sue letture e, successivamente, si è adottato un approccio comparativo sull'uso pausale in 10 interpretazioni di uno stesso testo a cura di 10 poeti contemporanei (5 M e 5 F, inclusa l'autrice)[1]. In alcuni indici del dato orale rosselliano analizzato potrebbero dunque essere ravvisati i riflessi di quanto si può ricostruire attraverso la disamina delle vicende cliniche testimoniate da Rosselli stessa in scritti come *Storia di una malattia* (1977) (Caputo, 2004; Giovannuzzi, 2012; Venturini & De March, 2010; Paris, 2020).

2. *«Storia di una malattia»: ricostruzione delle vicende cliniche rosselliane*

Come ricostruito dalla poetessa stessa, lo sviluppo delle sue patologie ha attraversato una serie di tappe significative e simboliche. Nel 1954, Rosselli fu sottoposta a Roma alla sua prima seduta di terapia elettroconvulsiva, seguita da vari shock insulinici. Questi trattamenti erano mirati a curare ciò che inizialmente sembrava un esaurimento nervoso o un disturbo del sonno. Nello stesso anno, le fu anche diagnosticata la schizofrenia paranoide, e da quel momento iniziarono numerosi anni di terapia psicanalitica, alternati a periodi di ricovero clinico. Una svolta cruciale nelle vicende cliniche di Rosselli si ebbe nel 1969, quando le fu diagnosticato il morbo di Parkinson, o più precisamente una lesione al sistema extrapiramidale. A seguito di questa diagnosi, iniziò a ricevere farmaci antispastici, che continuò ad assumere fino alla sua tragica morte per suicidio nel 1996. Alcuni critici letterari hanno sollevato dubbi sulla vera natura della sua malattia, ipotizzando che il morbo di Parkinson potesse essere stato utilizzato come una sorta di copertura per la sua schizofrenia paranoide, di cui ha sempre sofferto fin dalla prima adolescenza. In un quadro simile, l'assunzione regolare dei farmaci antispastici potrebbe quindi essere stata finalizzata a curare i possibili effetti dannosi lasciati dagli elettroshock sul suo sistema muscolare. Nonostante le difficoltà oggettive nel

[1] Riportiamo un elenco delle voci selezionate, a cui affianchiamo tra parentesi un'indicazione sulla tipologia di lettura con cui hanno riprodotto il componimento rosselliano, secondo la classificazione di Colonna (2022), basata nell'organizzazione prosodica in relazione al testo e così classificata: lettura metrica (M, fedele al verso), sintattica (SI, fedele alla punteggiatura e all'andamento logico-sintattico nel suo complesso), semantico-sintagmatica (SE; incentrata su unità sintagmatiche e semantiche del discorso poetico). Le cinque voci maschili sono quelle di A. Bertoni (lettura SE), M. Bordoni (lettura SI), F. Ottonello (lettura M), D. Piccini (lettura SI) e G.M. Villalta (lettura SE); quelle femminili, invece, di A.M. Carpi (lettura SE), I. Leardini (lettura M), F. Mancinelli (lettura M), R. Tempesta (lettura SI) e la stessa Rosselli. Per una descrizione dettagliata in termini stilistici di alcune delle voci selezionate cfr. Colonna 2021a.

ricostruire l'effettiva causa della sua potenziale disartria[2], l'analisi del suo modo di leggere ci offre interessanti informazioni sulla natura del suo eloquio e ci consente di valutare anche, all'interno del panorama italiano, i possibili riflessi prosodici di una condizione patologica (potenzialmente dovuta alla schizofrenia paranoide, al Parkinson o a una compresenza delle due). Infine, è particolarmente affascinante la poliedricità di questa autrice, che è in grado di conciliare la dimensione poetica con quella musicologica in uno stile plurilingue: è possibile riconoscere, infatti, una convergenza tra le sue strutture poetiche e le tecniche compositive della musica atonale a lei coeva (Rosselli, 1962).

3. *Metodologia e* corpus *delle letture selezionate*

Le letture selezionate per lo studio qualitativo sul parlato poetico di Rosselli sono 8 e sono le seguenti: *Figlia di un amore che ti divorò fui* (FI), *Impromptu* (IM), *La severa vita dei giustiziati rinnoverava* (LS), *Note che sorgono abissali* (NO), *Pietre tese nel bosco* (PI), *Rosa ripulita* (RO), *Se sinistramente* (SE), *Seguito dalle mosche, credendomi* (SM). A guidarci nella scelta delle registrazioni sono stati sia criteri di reperibilità del materiale, sia un'accettabile qualità del dato orale: cinque letture del *corpus* (FI, PI, LS, RO e SE[3]) provengono dalle *Teche Rai* e sono state interpretate da Rosselli in trasmissioni radiotelevisive; tre letture (NO e SE, lette in occasione del *Premio città di Recanati 1990*) sono, invece, presenti *online* e reperibili su YouTube; l'interpretazione di *Impromptu*, infine, è stata registrata e pubblicata come CD nel 1989, in occasione di una ristampa da parte della casa editrice genovese San Marco dei Giustiniani, nella collana «Quaderni di poesia». Le letture coprono un periodo di circa cinque anni – con il primo nucleo di registrazioni che risale al 1985 e l'ultimo al 1990 – e sono tratte da *Serie ospedaliera* (1969) e da *Documento* (1976), due delle maggiori raccolte della poetessa. Tra queste, come anticipato, abbiamo annotato integralmente l'interpretazione originale del poemetto *Impromptu* (1981), che – vista la sua particolare lunghezza, superiore ai quindici minuti – ci ha consentito di approfondire l'analisi della lettura e della voce di Rosselli, misurandola in una registrazione di un testo molto vario, ricco e complesso.

Dal punto di vista metrico-stilistico, i componimenti scelti mostrano una colorata variabilità: si va infatti da RO – la poesia più compatta del corpus, formata soltanto da 10 versi molto brevi e pressoché isosillabici (~ 8 sillabe per

[2] Come si riconosce dallo spoglio della letteratura sul tema, per Rosselli è complesso individuare confini precisi tra la schizofrenia, il Parkinson e le potenziali lesioni al sistema muscolare a causa degli elettroshock. In un simile quadro, le già ridotte pretese di associare la riconoscibile atipicità del parlato a un'unica causa diventano pressoché inesistenti. Pertanto, si avanza l'ipotesi che nella lettura rosselliana convivano i riflessi di tutte le condizioni descritte a più riprese.

[3] Le prime tre registrazioni sono provenienti da un gruppo di letture consecutive che Amelia Rosselli eseguì in occasione di una puntata della trasmissione radiofonica *Poeti al microfono*, condotta da Fabio Doplicher e Mario Mattia Giorgetti e in onda su Rai Radio 1 a partire dal gennaio del 1985. RO proviene direttamente dalle Teche Rai, mentre SE venne eseguita all'interno della trasmissione Rai, *L'Aquilone*, nella rubrica "Poeti in gara" (1989).

VS) – a IM – che invece è organizzato in tredici estese lasse e conta 355 versi, la cui lunghezza rimane però piuttosto regolare in tutto il poemetto (~ 10 sillabe per VS), nonostante globalmente ci siano escursioni sia verso 2 che verso 16 sillabe[4]. Tuttavia, come mostra dettagliatamente Lo Iacono (2023), tutti i testi rosselliani scelti sono stati composti da Rosselli rispondendo alle complesse e criptiche regole dello spazio metrico, la teoria filosofico-metricologica che prende il nome proprio dall'importante saggio in cui tale sistema viene messo a punto: *Spazi metrici* (1962). Rosselli non ricorre, dunque, né alle forme della metrica tradizionale, né al verso libero: opta, invece, per la sistematizzazione di un'architettura innovativa, i cui costituenti minimi metrici, per quanto flessibili e adattabili al variare delle geometrie, hanno anche dei confini regolari, che sono riconoscibili a uno sguardo e a un ascolto attenti. Questo, di conseguenza, rende la struttura (compositiva ed enunciativa) dei testi reciprocamente simile, a prescindere dalla lunghezza e dal numero dei versi (cfr. Lo Iacono 2023).

Anche in questa ricerca, come nei lavori di Colonna (2021a; 2022) e Lo Iacono (2023), il tipo predominante di parlato poetico è la lettura ad alta voce e non la recitazione a memoria. Questo viene confermato dalle fonti originali che abbiamo recuperato, dove possiamo vedere Amelia Rosselli leggere i suoi testi o sentire i conduttori radiotelevisivi che la presentano e introducono le sue letture. Inoltre, riguardo al registro linguistico su cui ci concentriamo, desideriamo precisare un aspetto già discusso da Colonna nella sua ricerca. La voce di Rosselli, così come tutte le altre voci presenti nell'archivio di VIP, sono mediate dal microfono e sono sempre accompagnate da riprese video o audio destinate a contesti specifici, come programmi radiofonici e televisivi. Come Colonna scrive, seguendo le osservazioni di Sessa (2018), "questo sicuramente influisce sul tipo di lettura, sulle diverse scelte adottate e sull'espressività specifica, che è anche legata alla gestualità" (Colonna, 2021a: 160). L'annotazione, come da protocollo dei lavori del progetto VIP (Colonna, 2021a; 2022; Lo Iacono, 2023), è avvenuta tramite l'applicativo *Praat* (Boersma, Weenink, 2021) e si è strutturata in 4 livelli, attraverso i quali si mette in luce la relazione tra testo e realizzazione prosodica: VS (Verso), EN (Enunciato poetico), CP (Curva Prosodica), PR (Parola ritmica). L'individuazione di EN e PR avviene su un piano percettivo: si definisce l'*Enunciato poetico* come "un atto linguistico indipendente, che presenti un confine terminale e una sua significazione prosodica unitaria, che può essere costituito da una sola o da più sezioni interne" (Colonna, 2021a: 163); mentre con *Parola ritmica* ci riferiamo non solo alle "unità tonali-accentuali" (Colonna, 2017: 92) su cui, nel *continuum* prosodico, ricade la cadenza ritmica (Colonna, 2021a: 176), ma anche "[al]l'insieme di una o più parole riprodotte foneticamente" in una sequenza con un unico accento forte (Ivi: 178). Infine, le *Curve prosodiche* sono unità interpausali, cioè unità prosodiche poste tra "porzioni di silenzio funzionali, che separano tra loro unità melodiche compatte"

[4] Riportiamo in elenco la struttura metrico-testuale degli altri componimenti: FI (19 vv; 13,7 sill/vs); LS (21 vv; 13,3 sill/vs); NO (23 vv; 12,3 sill/vs); PI (12 vv; 11,17 sill/vs); SE (37 vv; 9,6 sill/vs), SM (22 vv; 11 sill/vs).

(Colonna, 2021a: 166). Il riconoscimento delle CP avviene attraverso una valutazione percettiva che tiene in considerazione due precisi parametri, cioè «l'andamento di f_0 e dell'intensità (dB) (messi in evidenza sullo spettrogramma), compresi in una determinata unità di tempo» (Colonna, 2021a: ibidem). L'insieme delle analisi e dei grafici che utilizziamo a supporto delle descrizioni di ciascuna lettura è stato definito da Colonna *VIP-VIEW* (Colonna, 2021a; 2022). A comporre la *VIP-VIEW* sono grafici come il *VIP-Radar*, i *VIP-Histograms* e il *VIP-EnjambmentLine*, con cui conduciamo l'analisi qualitativa e descrittiva delle interpretazioni presenti nel nostro corpus, considerate individualmente. Il *VIP-Radar* include al suo interno una serie di parametri acustici e altri indici appositamente sviluppati, che presentiamo nella tabella (Tab. 1, riadattata da Colonna 2021b).

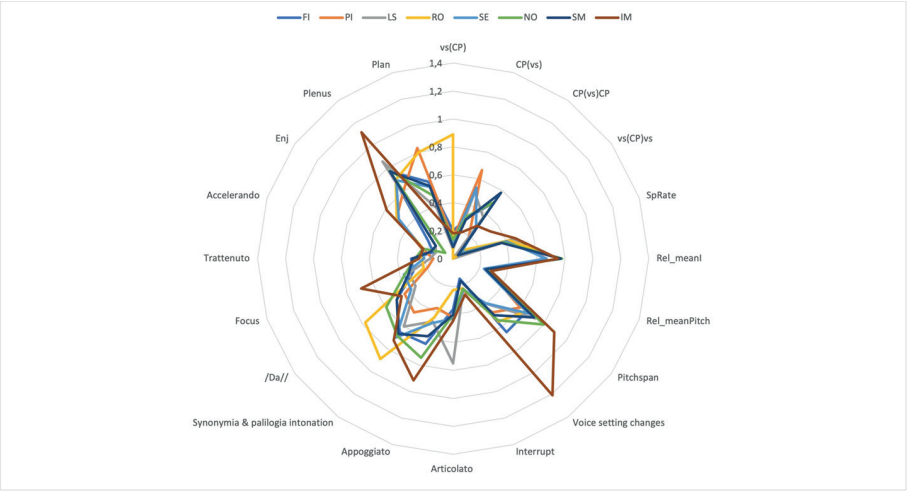
Tabella 1 - *I venti indici della metodologia VIP*

Indici	*Definizioni*
Versi curva vs(CP)	CP che coincidono con la linea del VS
Curve emiverso CP(vs)	CP che includono una porzione di VS
Curve interverso CP(vs)CP	CP che, tra due VV, includono una porzione di entrambi i VS
Curve bi-poliverso vs(CP)vs	CP che includono due o più VS per intero
Pitchspan	estensione melodica nell'intera lettura, calcolata in semitoni
Rel_meanpitch	frequenza media relativa (f_0) misurata in Hertz
Rel_meanI	intensità media relativa misurata in dB
Voice Setting Changes	numero di salti tonali o di registro in relazione al numero di CP
Speech Rate	velocità d'eloquio, data dal rapporto tra le sillabe fonetiche totali percepite e la durata totale delle CP di ogni lettura
Accelerando	accelerazione nella velocità elocutiva, definita a livello percettivo
Trattenuto	rallentamenti nella velocità elocutiva, definiti a livello percettivo
Plenus	rapporto tra durata totale di parlato (CP) e pause (P)
Focus	presenza di rilievi e prominenze focali all'interno delle letture
/Da//	intonazione *dichiarativa assertiva* e *poetica*, connotata da andamento non del tutto discendente, terminante su un livello medio-basso
Interrupt	frammentarietà della pronuncia, interno alle CP e nell'uso di P
Appoggiato	modalità accentuativa marcata dalle PR interne alle CP e risultato del rapporto tra numero di PR e CP
Articolato	scansione degli EN in CP, che implica un'intenzione di respirazione tra frasi. È ottenuto dal rapporto tra CP ed EN totali
Synonymia & palilogia	presenza di figure retoriche di ripetizione dell'intonazione delle CP, uguali e sullo stesso tono (*palilogia*) o uguali ma su toni diversi (*synonymia*), con possibilità di *variatio* (cfr. retorica musicale)
Enjambement	numero di *enjambement* realizzati a mezzo di pause
Plan	pianificazione del discorso poetico prosodico, ottenuta dal rapporto del totale degli EN sul totale dei VV

I primi quattro parametri – con cui descriviamo la relazione tra asse testuale e asse prosodico – rilevano la tipologia di CP rispetto al verso poetico. Tra gli altri indici relativi alle caratteristiche prosodiche si raggruppano quelli acustici e stilistici della

lettura poetica: frequenza media relativa (f_0), *pitchspan*, intensità media relativa (dB), *Voice Setting Changes*, *Speech Rate*, *Accelerando*, *Trattenuto*, *Plenus*, *Focus*, */Da//*, *Interrupt*, *Appoggiato* e *Articolato* (Colonna, 2021a; 2022; Lo Iacono, 2023). Attraverso l'indice di *Synonymia & Palilogia Intonation* individuiamo, invece, un altro aspetto rappresentativo della lettura poetica, ossia la fitta trama di riprese retoriche nell'intonazione. Infine, altre caratteristiche salienti rappresentate nel *VIP-Radar* sono la relazione tra la *mise en page* del testo poetico e la sua proiezione nella struttura prosodica (il numero di *enjambement* realizzati prosodicamente a mezzo di pause) o anche il grado di pianificazione del discorso poetico prosodico (indice di *Plan*). In Fig. 1 possiamo osservare il *General VIP-Radar* delle otto letture rosselliane considerate (cfr. Lo Iacono, 2023).

Figura 1 - General Vip-Radar *delle otto letture poetiche di Amelia Rosselli*



4. *La pausazione nella metodologia di VIP*

Per quanto riguarda l'individuazione e la classificazione delle pause, la metodologia del gruppo di ricerca VIP trova i suoi presupposti non solo nei lavori dei ricercatori del LFSAG (Romano, 2017; De Iacovo *et al.*, 2020), ma anche e soprattutto in alcuni dei riferimenti teorici più rilevanti sul tema come, per l'italiano, Bertinetto & Magno Caldognetto (1993). Molti altri sistemi classificatori per la pausazione sono sorti nel tempo, a partire dalla macro-distinzione tra pause vuote e pause piene (cfr. Mahl, 1956; Duez, 1982; Shriberg, 1994; Blankenship & Kay, 1964; o, per l'italiano, Giannini, 2001 e 2003), che per altro rappresenta uno dei punti teorici fondamentali nell'annotazione del discorso. È essenziale, ma al tempo stesso delicato, definire prima di tutto quando una pausa può essere considerata completamente vuota; esercizio che diventa problematico, ad es., quando si osservano fenomeni come il rallentamento delle vibrazioni delle corde vocali seguito

da una rapida ripresa, che fa sorgere il dubbio se classificarlo come un'interruzione o una transizione. In secondo luogo, è altrettanto complesso stabilire delle soglie qualitative, le quali, dipendendo dalla velocità del discorso, non solo distinguano tra pause motivate fisiologicamente o respiratorie e pause linguistiche intenzionali, ma identifichino anche le interruzioni legate a disfluenze o disturbi cognitivamente rilevanti. Nei tentativi di individuare e definire le soglie minime di riferimento per le pause non fluenti, i lavori di Goldman-Eisler (1958 e 1968) hanno svolto un ruolo chiave, stabilendo soglie minime di 200÷250 ms. Tuttavia, De Iacovo *et al.* (2020) evidenziano che in letteratura sono state proposte soglie minime sia più elevate sia più basse, basandosi su studi empirici[5], anche se è importante sottolineare che la definizione di una soglia che non tenga conto della variabilità individuale e delle normali accelerazioni e decelerazioni del discorso rischia di introdurre distorsioni nelle valutazioni della produzione orale di un individuo. Nella metodologia di VIP sono prese in considerazione e quindi annotate esclusivamente queste pause, che – più o meno brevi che siano – non interrompono il flusso logico del discorso e permettono che l'enunciato continui dopo la pausa. A caratterizzare le pause piene – che vengono da noi commentate soltanto in descrizioni qualitative – sono, invece, sia i «fenomeni interiettivi (trascritti con <eeh> per la semplice vocalizzazione, <ehm> per quella con nasalizzazione)», sia gli «allungamenti dell'ultima vocale o consonante di parola, marcate con la duplice ripetizione della vocale <vv> o consonante <cc> alla fine dell'elemento lessicale interessato, indipendentemente dal timbro e durata effettivi» (Romano, 2008: 187). Altre proposte di classificazione di pause piene e disfluenze sono state proposte, ad es., dai lavori di Blankenship & Kay (1964), di Shriberg (1994), Lickley (1994), Heeman & Allen (1999) ed Eklund (2004), i cui studi convergono nell'individuazione di cinque disfluenze (Pause piene, Ripetizioni, Sostituzioni, Inserzioni, Cancellazioni).

Il registro del parlato poetico – che ha richiesto una considerazione specifica fin dalle prime analisi – ha fatto percepire tuttavia la necessità di sviluppare una terminologia specifica anche per quanto riguarda la pausazione. Così, partendo dall'impianto teorico delineato da Romano (2017) per il parlato spontaneo e radiofonico, da una parte, è stato ampliato il ventaglio delle possibili sfumature di lunghezza delle pause classificate; e dall'altra, è stato proposto l'inserimento di un'ulteriore tipologia di pausa intermedia: la <pm>. Chiaramente, in uno studio condotto su un registro di parlato fortemente connesso anche alla struttura testuale è necessario considerare la sia netta distinzione tra *timing-based pauses* e *planning-based pauses*, cioè tra pause obbligatorie per la struttura dell'enunciato e del testo e pause prodotte per rendere più complessa la struttura testuale pianificata (Colonna, 2021a: 175). Alla luce di quanto detto, le etichette utilizzate comprendono pause

[5] Altri autori (O'Shaughnessy, 1992) riconoscono che tale soglia potrebbe aumentare per le pause ai confini sintattici principali e stabiliscono valori intorno a 490÷500 ms. Nonostante esista un accordo sostanziale su queste misurazioni, ci sono studi che utilizzano soglie più brevi. A titolo di esempio, come sottolineato da Ciaurelli (2020), Eklund (2004) e Kendall (2009) hanno preferito assegnare valori di 60÷90 ms per stabilire minimi naturali.

brevi <pb> (>0,1 s, <0,4 s); pause medie <pm> (>0,4 s, <0,6 s); pause lunghe <pl> (>0,6 s, <1 s); pause molto lunghe <pll> (>1 s); e, alla fine di una strofa, pause reset <P>. L'inventario di queste pause è stato individuato con un'alta frequenza in diversi parlanti, confermando così le misurazioni proposte inizialmente, anche se alcuni di essi le utilizzano solo in modo limitato. (Colonna, 2021a: 175-176). La variazione del ventaglio delle pause in alcune letture può dipendere da molteplici fattori, sia legati alle esigenze stilistiche dei poeti sia alla struttura del testo scritto, sul quale corre – a velocità diverse – l'interpretazione sintattica da un lato e l'interpretazione metrica dall'altro (Bertoni, 2006). Riguardo alla pausazione e all'interpretazione metrica, si fa riferimento soprattutto agli scritti di Menichetti (1993), che ha proposto una classificazione includente la pausa vera e propria associata a un'intonazione specifica (//); la giuntura sintagmatica come una pausa virtuale non percepibile (/); e l'intonazione sospensiva o d'attesa, caratterizzata da rallentamento e interruzione, suggerendo inoltre che la pausa della fine del verso non sarebbe prosodicamente una vera pausa, quanto piuttosto una figura ritmosillabica, realizzata solo nella metrica tradizionale. Alla luce dei protocolli di annotazione in uso nel progetto CLIPS (cfr. Savy, 2006; Savy, 2007) e degli studi condotti presso il LFSAG (cf. Romano, 2008; De Iacovo *et al.*, 2020), oltre alle etichette presenti nella metodologia del progetto VIP descritte prima, per sviluppare la parte di studio comparativa e soffermarci più dettagliatamente sulla dimensione pausale del parlato poetico, sono stati indagati ulteriori parametri salienti, i quali hanno avuto dei riflessi nel sistema di annotazione: sono state registrate difatti anche le pause con glottalizzazione (indicata, a seconda della sua posizione, rispettivamente: <?pbb>; <pbb?>; <?pbb?>) e le inspirazioni (<insp>).

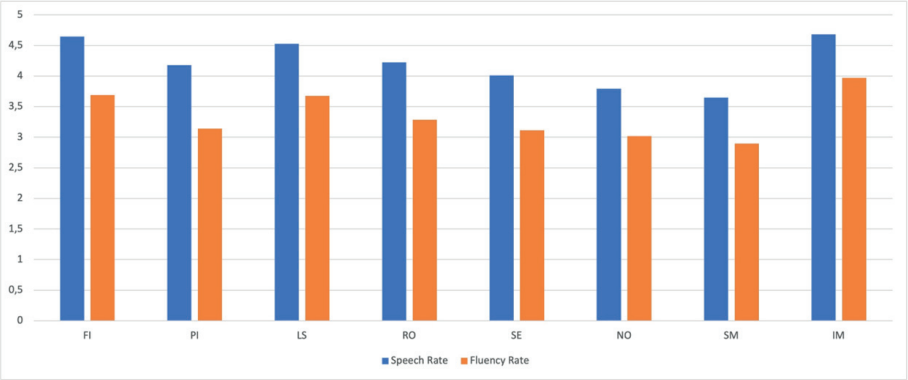
5. *Melodia e silenzio: lo stile prosodico di Amelia Rosselli*

Effettueremo una breve descrizione dello stile rosselliano utilizzando alcuni grafici, che, insieme al *General VIP-Radar*, ci consentiranno di confrontare i dati emersi dalle 8 letture analizzate: ricorreremo allo *Speech-FluencyRate-VIP-Histogram*, relativo a *Speech Rate* e *Fluency Rate*[6]. Successivamente, mostreremo il *CP&P-VIP- Histogram*, che cattura il rapporto tra le medie delle durate totali di CP e P e l'*Appoggiato-articolato-VIP-Histogram*, che riporta, confrontandoli, il grado di *appoggiato* e *articolato*. Oltre a questi ci avvarremo dell'*Enjambement-VIP-Histogram*, in cui vengono riportate per ogni lettura le percentuali di *enjambement* riprodotti con pausa, di *enjambement* riprodotti senza pausa e di allungamenti compensativi. Infine, attraverso il *VIP-CPVS-BoxPlot* mostreremo nel complesso le scelte prosodico-organizzative alle quali Rosselli ricorre nelle otto letture[7].

[6] Come indicato in Tab. 1, lo *Speech Rate* misura la velocità elocutiva in base al rapporto tra sillabe totali e durata di CP, mentre il *Fluency Rate* descrive la scioltezza, considerando le sillabe totali e la durata totale della lettura, comprensiva di pause (cfr. Colonna, 2022).

7 I dati, le analisi e i relativi grafici di questa sezione sono tratti e riadattati da Lo Iacono (2023).

Figura 2 - *Speech-FluencyRate-VIP-Histogram* comparativo

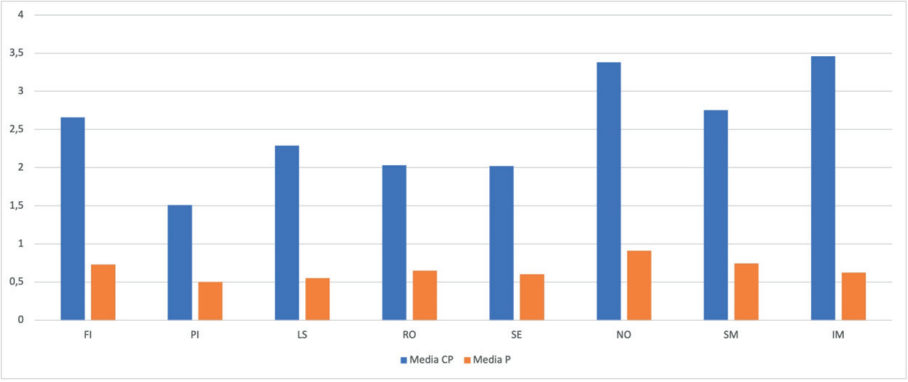


Dall'analisi del grafico (Fig. 2), si osserva una certa variabilità tra le otto letture annotate: chiaro risultato del fatto che Rosselli prevedeva e concepiva per ogni testo una struttura unica e particolare, con un ordine interno sempre diverso. Di conseguenza, i componimenti con una struttura simile, come ad es. RO e PI, o FI e LS, o ancora NO e SM, tendono ad avere valori simili. Inoltre, osservando i valori della deviazione standard di un altro parametro della nostra *VIP-View*, la *CP-Speech Rate*, si nota che Rosselli mantiene tendenzialmente la stessa velocità di eloquio durante tutte le interpretazioni. Ciò conferma quanto già suggerito da Colonna (2021a, 2022), ovvero che Rosselli cerca realmente "una sorta di inseguimento formulare di qualcosa che si cerca e ripete" in modo regolare, ordinato e ricorsivo (Colonna, 2021a: 436). È interessante notare che questo elemento spazio-temporale che si ripete in modo ridondante non è tanto il verso, come sosteneva Rosselli, ma piuttosto un'unità logica e metrico-prosodica più vicina a ciò che nella metodologia di VIP chiamiamo CP. Proseguendo ancora con la descrizione del grafico (Fig. 2), è possibile osservare una fascia bassa (<4 sill/s) e una fascia medio-alta (suddivisibile ulteriormente in due sottogruppi, il primo con velocità compresa tra ≥4 e <4,5 sill/s e il secondo con velocità ≥4,5 sill/s): alla prima appartengono le *Speech Rate* di NO e SM (la più bassa); al secondo gruppo appartengono le *Speech Rate* medie di PI, RO e SE e le medio-alte di FI, LS e del poemetto IM (la più alta).

Analizzando, invece, i dati di *Speech Rate* e *Fluency Rate*, si osserva che i loro andamenti non procedono in modo parallelo e differiscono per ciascuna lettura, confermando quanto già rilevato da Colonna, la quale sostiene che il rapporto tra *Speech* e *Fluency Rate* serve a dimostrare che "indipendentemente dalla velocità d'eloquio interna alle curve prosodiche, l'impiego delle pause sia molto divergente, influenzando anche una diversa percezione" (Colonna, 2021a: 439). Si nota, inoltre, un divario massimo tra i valori di *Speech Rate* e *Fluency Rate* nella lettura di PI, in cui l'indice di *Fluency Rate* risulta inferiore di 1,04 rispetto a quello di *Speech Rate*: ciò è motivato dal bilanciamento della durata melodica con la durata complessiva delle pause. Tuttavia, in generale, i range tra *Speech Rate* e *Fluency Rate* nelle letture di Amelia Rosselli sono piuttosto ampi, confermando quanto

affermato da Colonna riguardo ai poeti sperimentali: «[l]a parte invece che ingloba le voci "sperimentali", oltre a mantenere più alta la media di *Speech* e *Fluency Rate*, mantiene alta, come avevamo ipotizzato, anche la variazione interna (si raggiungono i massimi di deviazione standard)» (Colonna, 2021a: 439). Infine, possiamo aggiungere che la velocità d'eloquio media di Rosselli è di circa 4,2 sill/s, un valore medio all'interno della metodologia di VIP. L'indice di scioltezza media, invece, si attesta intorno a 3,35 sill./s. Descriviamo ora il *CP&P-VIP-H* (Fig. 3), che cattura i valori delle durate medie delle Curve prosodiche e delle pause delle otto letture prese in considerazione. Riguardo alla durata delle CP, confrontando le varie interpretazioni, è possibile riconoscere, da una parte, una discreta variazione interna alle letture e, dall'altra, dei potenziali raggruppamenti tra i componimenti simili tra loro formalmente: ad es. le già citate RO e PI; o ancora FI e LS. Inoltre, dalle analisi comparative non è emersa alcuna correlazione tra la durata media delle CP e il numero medio di sillabe presenti nei versi delle poesie: questo risultato sottolinea ulteriormente la profonda differenza tra le unità prosodico-concettuali e le unità versali nel sistema metrico di Rosselli.

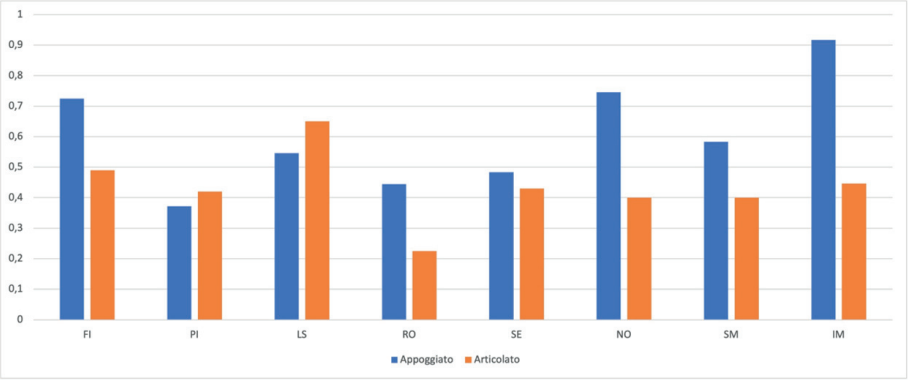
Figura 3 - *CP&P-VIP-Histogram* comparativo



Analizzando la durata media delle pause, si conferma la presenza di variabilità nelle letture. Dal grafico, inoltre, non emerge una correlazione né diretta né inversa tra la durata delle pause e la durata delle CP. Questo risultato corrobora le analisi condotte da Colonna su Rosselli (Colonna, 2021a: 436-437), evidenziando, da un lato, un *modus legendi* che si allinea con le altre voci sperimentali e la loro variabilità nell'alternanza tra pause ed eloquio, e, dall'altro, rafforza l'idea che la pausa non sia una semplice conseguenza della lunghezza delle CP, ma piuttosto un principio costruttivo della metrica rosselliana, indipendente e necessario tanto quanto l'organizzazione dei costituenti prosodici. Nel complesso, la durata media delle CP si attesta intorno a 2,5 s, un valore che può essere considerato medio-alto; anche la durata media delle pause, pari a 0,65 s, può essere considerata relativamente alta. Analizzando i valori delle singole letture, emergono ulteriori aspetti rilevanti: si nota, ad es., una tendenza alla maggiore durata delle curve

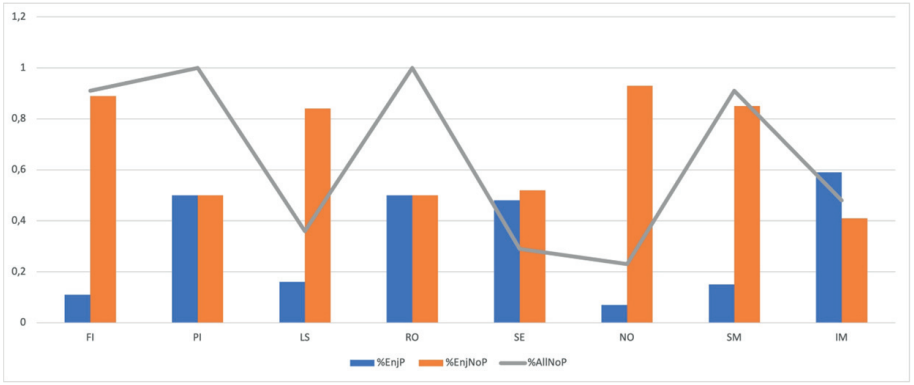
prosodiche nelle interpretazioni tratte dalla raccolta *Documento* rispetto a quelle di *Serie ospedaliera*. Non trascurabile è anche la marcata lunghezza delle CP nella lettura di IM, che presenta il valore medio più elevato dell'intero corpus (oltre 3,5 s). In particolare, nella lettura di IM si può osservare una chiara sovrapposizione tra la struttura metrica (in cui spicca la terzina come elemento metrico predominante) e la struttura prosodica della lettura: questo mette in luce la corrispondenza tra le unità di pensiero logico-sintattiche, che Rosselli analizza dettagliatamente in *Spazi metrici* (1962), e le lunghe unità metrico-prosodiche della metricologia rosselliana.

Figura 4 - *Appoggiato-Articolato-VIP-Histogram* comparativo



Attraverso l'analisi dell'*Appoggiato-articolato-VIP-Histogram* (Fig. 4), possiamo esaminare e descrivere le relazioni che legano le parole ritmiche alle curve prosodiche (*appoggiato*) e le curve prosodiche agli enunciati (*articolato*). Complessivamente, nel modo in cui Rosselli legge i testi, come evidenziato dal grafico, prevale un profilo *appoggiato* (con una media di circa 0,6) rispetto a quello *articolato* (con una media di circa 0,45): ciò indica che a essere marcate sono dunque le PR, le unità minori. Solo in due delle otto letture prevale l'*articolazione*, seppur con una differenza minima (PI e LS, che ha il valore di *articolato* più alto dell'intero *corpus*, pari a 0,65). In tutti gli altri casi, prevale nettamente l'*Appoggiato*, con valori molto più alti e in particolare, spicca l'alto valore dell'indice di *Appoggiato* nell'interpretazione di *Impromptu*, che si avvicina a 0,9. I dati confermano la tendenza individuata da Colonna, secondo cui nelle voci sperimentali l'*appoggiatura* è marcatamente più presente, pur mostrando una considerevole var. st. e quindi grande variabilità interna (Colonna, 2021a: 443).

Figura 5 - *VIP-Enjambment-Histogram* comparativo

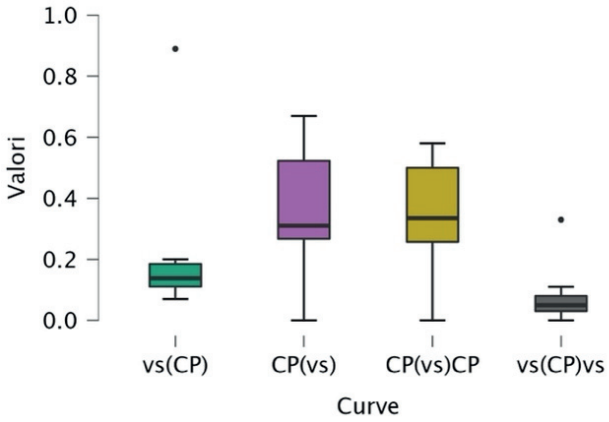


Concludiamo la descrizione dei *VIP-Histograms* presentando il *VIP-Enjambment-Histogram* (Fig. 5) di cui ci serviamo per riportare le percentuali delle tipologie di inarcature (EnjP ed EnjNoP) presenti rispetto al totale e la quantità di EnjNoP soggetti ad allungamenti compensativi. Se escludiamo la lettura di *Impromptu* – la cui struttura metrica molto varia e complessa inevitabilmente influisce nella lettura – nello stile di lettura di Rosselli prevalgono gli *enjambement* riprodotti senza ricorrere a pause tra *rejet* e *contre-rejet*. Ciò concorre, ascoltando le letture di Rosselli, a rafforzare una sensazione già trasmessa anche dall'osservazione del testo grafico stesso, ossia la consapevolezza di fare esperienza di componimenti compatti, ordinatamente unitari e *cubici*[8]. Inoltre, questo fluire continuo viene enfatizzato anche attraverso i molti allungamenti compensativi a cui la poetessa spesso ricorre nel riprodurre le inarcature prive di pausa: in quattro delle otto letture – come mostra la linea grigia nel grafico – la percentuale di AllNoP rispetto al totale di EnjNoP è addirittura superiore al 90 %. Nell'analizzare il comportamento di Rosselli in occasione delle inarcature, tuttavia, è emerso anche un ulteriore aspetto interessante, ossia il ricorso agli *pseudo-enjambement*. Queste, considerabili come normali EnjNoP, sono tuttavia inarcature interessate da una sorta di rapporto di tensione e distensione tra *rejet* e *contre-rejet*, dal momento che la pausa che normalmente le separa non è affatto assente, bensì soltanto anticipata in una posizione di *pre-rejet* o posticipata in una posizione immediatamente successiva al *contre-rejet*. Sebbene semanticamente e prosodicamente queste inarcature abbiano effetti simili

[8] Per quanto nelle sue innumerevoli autoanalisi Rosselli abbia più volte fatto ricorso (consapevolmente o meno) a una terminologia instabile, l'accostamento formale dei suoi componimenti alle figure geometriche del quadrato o del cubo è un vero *leitmotiv*. Come viene infatti riportato in letteratura (cfr. tra gli altri Manera, 2006), lo *spazio metrico* rosselliano consta di monadi ritmiche dalle geometrie variabili, che a più livelli della lingua richiamano forme compatte e geometriche. Tutte queste forme metrico-testuali regolari, sempre secondo Rosselli, nel momento dell'enunciazione poetica vengono proiettate nella dimensione temporale ed è proprio attraverso la sovrapposizione tra quest'ultima e la dimensione spazio-visuale che il quadrato poetico si fa cubo dinamico ed energetico (Lo Iacono, 2023).

agli *enjambement* forti, in realtà contribuiscono a sottolineare la sensazione di sfasamento tra la struttura grafico-formale e la struttura metrico-sintattica. Inoltre, rafforzano la necessità del lettore di scorrere continuamente da un verso all'altro, facendo emergere un senso di unità soltanto alla fine del componimento. Quest'uso degli *pseudo-enjambement* enfatizza la complessità della relazione tra forma grafica e struttura metrico-sintattica, evidenziando la tensione presente nel processo di lettura e sottolineando la peculiarità dell'approccio di Rosselli alla composizione poetica. Visto il latente brulichio di ritmi a cui la poetessa ricorre, ascoltando Rosselli leggere non è facile classificare quello che Manera definisce «il movimento intonativo-enunciativo» (Manera, 2006: 43). Tuttavia, è, invece, molto facile riconoscere e apprezzare quanto le sue scelte prosodiche e – di conseguenza – la struttura delle CP e degli EN tendano a far affiorare un maggiore rispetto per le unità logiche piuttosto che per quelle metrico-grafiche. Rosselli, anche in questo caso, riconferma nella pratica di lettura quanto condensa nei suoi scritti musico-metricologici: l'unità versale viene riprodotta rispettandone i confini metrici solo quando è sovrapponibile e sovrapposta all'unità logica di pensiero. In quasi tutti gli altri casi tra i confini metrici dei versi e i confini prosodici delle CP non v'è simmetria e prevale così una gestione che definiamo logico-sintattica.

Figura 6 - *VIP-CPVS-Boxplot* comparativo



Il comportamento e le scelte stilistico-prosodiche di Rosselli possono essere mostrati attraverso il nostro VIP-CPVS-Boxplot (Fig. 6), in cui confluiscono i dati delle otto letture del corpus: osserviamo rispettivamente i valori e le distribuzioni dei versi-curva vs(CP), delle curve emiverso CP(vs), delle curve interverso CP(vs)CP e delle curve bi-poliverso vs(CP)vs. Dall'analisi del grafico emerge una tendenziale prevalenza di curve emiverso CP(vs): il 50 % delle letture ne contiene, infatti, un numero che oscilla tra il 27 % e il 52 % del totale delle CP; mentre un quarto delle letture selezionate ne contiene addirittura oltre al 60 %, con un massimo di 0,67. Meno frequenti sono, invece, i *versi-curva* e le *curve bi-poliverso*: per quanto riguarda i primi, è importante notare che in tutte le letture di Rosselli ne viene riprodotto

almeno uno, come indicato dal fatto che il valore minimo nel rispettivo *box* non è zero. La distribuzione dei dati per i *versi-curva*, inoltre, è molto concentrata in una fascia ristretta, in quanto il 50 % delle letture contiene un numero di vs(CP) che rientra tra l'11 e il 18,5 % del totale delle CP; a questa distribuzione si accompagnano un estremo superiore non *outlier* che non va oltre il 19 % e un estremo inferiore non *outlier* che non scende sotto il 12 %. Per quanto riguarda le curve bi-poliverso, possiamo rilevare che sono la tipologia di CP statisticamente meno presente nel corpus delle letture: nel 50 % dei casi la percentuale delle vs(CP) vs oscilla tra il 3 e il 8 %, con una media simile alla mediana e una distribuzione tendenzialmente simmetrica. Nel complesso, quindi, possiamo riconoscere una decisa tendenza di Rosselli a prediligere una lettura che con gli strumenti della metodologia di VIP definiremmo logico-sintattica: scelta stilistica che conferma la tendenza della poetessa a creare un legame stretto tra la struttura logico-sintattica e quella prosodica dei suoi componimenti.

Per concludere la descrizione dello stile di lettura rosselliano, esaminiamo gli indici del nostro *General VIP-Radar*. Partendo da un'osservazione globale, possiamo notare a prima vista sia concentrazioni di valori piuttosto significative per alcuni parametri (stilistici, ma anche prosodico-testuali), sia letture con valori estremamente diversi rispetto alle altre, confermando i tratti comuni e le divergenze colte dagli altri grafici. Due esempi emblematici sono le due letture outlier nel nostro corpus: RO e IM. La prima si distingue su tutte, da una parte, per la grande presenza di versi curva e, dall'altra, per gli alti valori degli indici di *Synonymia & Palilogia Intonation* e di */Da//*: è, quindi, la lettura più metrica ed è anche quella con la struttura retorica più ridondante del corpus. La seconda, invece, spicca per l'alto tasso di *Voice Setting Changes*, di *Focus*, di *Appoggiato* e per il numero di *enjambement* forti (EnjP); tra le otto è, infatti, la lettura in cui Rosselli usa lo strumento vocale nel modo più sperimentale e vario, portando all'estremo una tendenza che però è presente anche nelle altre interpretazioni, visto che sono tutte melodicamente molto colorate da sbalzi tonali evidenti e cambiamenti di registro vocale. Osserviamo, a tal riguardo, che IM è anche la lettura con il più alto indice di *Pitchspan*, nonostante i valori delle interpretazioni siano decisamente più concentrati: l'utilizzo di un ampio ventaglio di tonalità è, infatti, una delle caratteristiche più evidenti dell'interpretazione orale del poemetto e, più in generale, anche dello stile di lettura di Rosselli, la cui voce diventa lo strumento melodico attraverso cui riprodurre quel formicolio ritmico con cui la poetessa cercava di riempire il suo spazio metrico. Esaminando i parametri in cui è riscontrabile una maggiore convergenza dei valori, innanzitutto notiamo – e di conseguenza confermiamo – una netta prevalenza di *curve emiverso* e di *curve interverso*, con due picchi evidenti: le letture di PI e di SM. Nel complesso, la velocità d'eloquio rimane stabile, con la maggior parte delle letture che si attesta in un *range* ristretto: quella di SM è la lettura che presenta il tasso di *Speech Rate* più basso (~0,36), mentre quello più alto (~0,47) si registra nella lettura di IM.

Non inaspettatamente, inoltre, possiamo notare una decisa convergenza dei dati riguardo ai parametri stilistici e ai parametri della qualità della voce (frequenza

media relativa, intensità media relativa e intervallo tra il tono più alto e quello più basso): se escludiamo il caso già discusso di IM e del suo tasso di *Pitchspan*, rileviamo che quasi tutti gli altri valori sono ravvicinati nel *General VIP-Radar*. Ciononostante, analizzando il radar e ascoltando le interpretazioni è possibile non solo rilevare e apprezzare le oscillazioni della frequenza media relativa, ma è possibile anche motivarle, confrontandone i valori con il contenuto dei testi. Riconosciamo, in questo modo, una tendenza (che potremmo definire fonosimbolica) di Rosselli a servirsi di una frequenza fondamentale mediamente più bassa nella lettura di testi dai contenuti più cupi, tragici e drammatici, come in SM in cui ritornano, assillanti, il tema della persecuzione e quello della paranoia. Per contro, nei componimenti invece più dinamici, energici e variegati – come IM – rileviamo la tendenza della poetessa ad aumentare la frequenza fondamentale media. Proseguendo con la descrizione dei parametri stilistici, anche circa l'indice di *Pitchspan* possiamo riconoscere una discreta convergenza dei dati delle letture verso valori relativamente alti, peculiarità di un *modus legendi* decisamente sperimentale e complesso, in cui la struttura melodico-musicale acquista e mantiene un valore strutturante. Se associamo a questa osservazione anche l'analisi dei parametri melodico-retorici, ci rendiamo conto – confermando ulteriormente le intenzioni teoriche della poetessa – dell'estrema importanza del sostrato musicale e melodico delle liriche: una centralità che abita l'atto di creazione dei testi e che si rinnova continuamente, vivendo lo spazio metrico in ogni sua dimensione, anche in quella orale. Infatti, nella lettura l'alto tasso dell'indice di *Synonymia & Palilogia Intonation* – accompagnato dalla presenza costante di un caratteristico profilo di *Interrupt* – è la conseguenza diretta di due aspetti cruciali nella teoria metricologica rosselliana. Da una parte, la sua affinità con il serialismo dodecafonico, con il webernismo, il post-webernismo e con la musica elettronica (Dolce, 1988: 37-45), e dall'altra, la centralità che hanno sia il concetto di ridondanza, sia la profondità della struttura retorica. Tra i parametri più interessanti del nostro grafico v'è senza dubbio quello del *Focus*, attraverso il quale siamo in grado di individuare i casi in cui Rosselli, ricorrendo a delle prominenze focali, ha rivestito di particolare importanza alcuni passi dei testi.

Nel complesso, possiamo notare una convergenza dei dati nel *General VIP-Radar*, ad eccezione dell'outlier IM: in quasi tutte le altre letture l'indice oscilla tra lo 0,2 e lo 0,4. L'aspetto decisamente interessante, tuttavia, risiede nell'aver rilevato una tendenza di Rosselli a focalizzare l'ultimo sintagma o direttamente l'ultima parola dei versi (e delle CP), a prescindere dalla categoria grammaticale delle parole coinvolte. Pertanto, congiunzioni, preposizioni, avverbi di negazione, articoli e pronomi sono spesso enfatizzati con dei rilievi focali che finiscono per semantizzare o risemantizzare questi elementi linguistici, inevitabilmente. Il *General VIP-Radar* mostra, inoltre, una tendenziale convergenza anche per quanto riguarda due parametri agogici della metodologia, ossia l'indice di *Accelerando* e di *Trattenuto*. Entrambi sono relativamente bassi nelle letture rosselliane e ciò è sintomatico di uno stile in cui ogni elemento della struttura poetica ha uno spazio e un'importanza uguali nella struttura prosodica, così come nell'impaginazione grafica, dove l'ordine

perfettamente regolare era garantito dalla scrittura a macchina. Per quanto riguarda gli ultimi tre parametri del *VIP-Radar* comparativo, possiamo ribadire la tendenza di Rosselli a riprodurre gli *enjambement* senza ricorrere all'uso di pause, preferendo una lettura più fluida che unisce *rejet* e *contre-rejet*. Anche in questo caso, il valore più alto è quello dell'interpretazione di IM, con circa il 60 % delle inarcature riprodotto come EnjP. Concludendo la descrizione del *General VIP-Radar*, anche i valori degli indici di *Plan* e di *Plenus* presentano alcune caratteristiche interessanti, veicolando informazioni decisive. I valori dell'indice di *Plenus*, ad es., confermano il peculiare uso delle pause e del silenzio: questo – accompagnato al costante indice di *Interrupt* – non può che connotare lo stile di lettura rosselliano come inconfondibilmente frammentario. Soffermandoci, infine, sull'indice della pianificazione del discorso poetico e valutando il comportamento di Rosselli nelle letture ci si accorge di una maggiore variazione nei dati: da un minimo di 0,27 di IM si arriva al massimo di 0,83 di RO. Il *General VIP-Radar* permette, dunque, di constatare un'ampia variazione tra le letture, con un netto scostamento di un caso rispetto agli altri (IM); al contempo, è tuttavia possibile isolare diversi punti di convergenza nelle scelte sonore, stilistiche e organizzative.

6. *La pausazione nel* modus legendi *di Amelia Rosselli:* tra fattori stilistici e riflessi di parlato patologico

Dall'analisi qualitativa sulla voce di Rosselli emergono dati interessanti riguardo alla particolare pausazione della poetessa e all'importanza dei silenzi nel suo stile di lettura. Concentrandoci sul ventaglio e sulla tipologia di pause utilizzate, i risultati ottenuti dalle ricerche (Fig. 7 e Tab. 2) mostrano un frequente utilizzo di colpi di glottide e pause brevi, alternati a pause lunghe e pause molto lunghe. Nel dettaglio, le <pb> mostrano una durata media che oscilla intorno ai 350 ms e una dev. st. media di ~0,137; le <pl>, invece, si attestano con una media superiore a 1000 ms e una dev. st. media di ~0,5. Questi dati sono una conferma dell'affidabilità del sistema di annotazione di VIP, con una soglia che, intorno ai 400 ms, segna uno scollamento nell'inventario pausale di Rosselli. La non trascurabile frequenza dei colpi di glottide su <pbb>, che abbiamo assimilato in questa quantificazione al calcolo delle <pb>, così come l'ulteriore presenza di <insp>, da una parte, si caratterizzano come tratto stilistico e, dall'altra, lasciano aperta un'ipotesi di riflesso di una atipicità presente, a maggior ragione se confrontate con i dati emersi dall'analisi della pausazione nelle letture di altri lettori dei testi rosselliani.

Figura 7 - *Valori delle pause nelle interpretazioni di Amelia Rosselli*

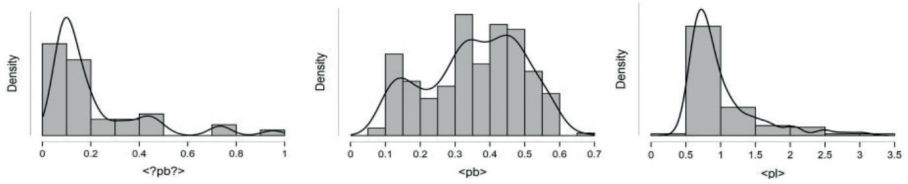


Tabella 2 - *Tabella descrittiva delle pause nelle interpretazioni di Amelia Rosselli*

<p>	*n.*	*Media*	*Varianza*	*Dev. st.*	*Min.*	*25 % q.*	*Mediana*	*75 % q.*	*Max.*
<?pb?>	44	0,213	0,041	0,202	0,062	0.090	0.120	0,233	0,952
<pb>	303	0,355	0,019	0,137	0,095	0,262	0,366	0,462	0,665
<pl>	97	1,004	0,255	0,505	0,477	0,680	0,824	1,131	3,047

Osservando, infatti, il grafico in Fig. 8 e la relativa tabella descrittiva (Tab. 2), possiamo riconoscere a colpo d'occhio una divergenza dei dati rispetto ai valori della pausazione nelle letture di Rosselli: situazione che ribadisce e conferma l'originalità prosodica della poetessa. I grafici, in questo caso, riportano i valori della pausazione ottenuti dall'analisi comparativa delle letture di dieci poeti (tra i quali è presente anche Rosselli) che si sono cimentati nell'interpretazione di *Pietre tese nel bosco*, una lirica presente nel *corpus* selezionato anche per lo studio qualitativo. Come possiamo osservare, per quanto le <pb> dei dieci interpreti risultino simili a quelle del grafico precedente – con una media di 385 ms e una dev. st. di 0,138 – concentrandoci sulle <pl> si nota, invece, una durata media complessivamente inferiore: 860 ms, con una dev. st. di 0,27. In aggiunta, è importante soffermarsi su un ulteriore dettaglio: la quantità di glottalizzazioni rilevate è, infatti, decisamente più bassa e non determinante (solo 9 casi individuati, contro le 44 <pb> con glottalizzazione rilevate nell'analisi del parlato poetico rosselliano). Ciò non solo ha motivato la scelta di non riportare la distribuzione delle pause con glottalizzazione nella Fig. 8, ma ci ha permesso anche di constatare una netta divergenza rispetto al *modus legendi* rosselliano, che invece è proprio caratterizzato, come abbiamo visto, da colpi di glottide che precedono e seguono le pause brevi.

Figura 8 - *Valori delle pause nelle interpretazioni di 10 poeti, inclusa Amelia Rosselli*

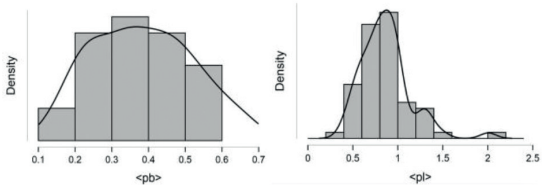


Tabella 3 - *Tabella descrittiva delle pause nelle interpretazioni di 10 poeti, inclusa Amelia Rosselli*

<p>	*n.*	*Media*	*Varianza*	*Dev. st.*	*Min.*	*25 % q.*	*Mediana*	*75 % q.*	*Max.*
<?pb?>	9	0,129	0,002	0,040	0,102	0.106	0.112	0,126	0,224
<pb>	89	0,385	0,019	0,138	0,129	0,259	0,379	0,480	0,694
<pl>	63	0,862	0,077	0,277	0,386	0,680	0,846	0,979	2,006

Circa la presenza della lettura di Rosselli tra le 10 interpretazioni selezionate è importante fare una precisazione: questa – utile per avere più dati a disposizione – non solo non ha alterato la valutazione analitica congiunta con le altre 9 letture, ma ne ha rafforzato l'efficacia. La maggior parte delle <?pb?> rilevate dall'analisi (cfr. Fig. 8) sono state, infatti, riscontrate nella lettura di Rosselli, permettendoci due importanti spunti di riflessione. Da una parte, innanzitutto, è stato possibile riconoscere che complessivamente le letture degli altri nove poeti sono scarsamente interessate dal caratteristico uso rosselliano dei colpi di glottide (che precedono e seguono le <pbb>), il quale rappresenta – all'interno di questa analisi congiunta con altre interpretazioni – un caso unico che non consente una vera comparazione su tale misura. Dall'altra, infine, è stato possibile confermare che la presenza di glottalizzazioni nel *modus legendi* rosselliano è una caratteristica pervasiva, da approfondire e motivare.

L'ipotesi che avanziamo alla luce delle nostre analisi è che la grande variabilità del ventaglio pausale rosselliano e la rilevante presenza di glottalizzazioni siano fenomeni che non rispondono soltanto a esigenze stilistiche e potrebbero mostrarsi, infatti, anche come possibili riflessi di un sostrato patologico. In un quadro generale che resta assai difficile da decifrare, tuttavia, la sensazione è che allo stile di lettura si associno anche elementi che sfuggono al controllo diretto della poetessa e che tradiscono una condizione di forte atipicità. Per quanto riguarda l'alto tasso di glottalizzazioni, ad es., a permettere di disambiguarne la natura concorrono, innanzitutto, l'assenza di una correlazione regolare con precisi *loci* prosodici, e, conseguentemente, l'impossibilità di riconoscerne una distribuzione motivata da fattori metrico-prosodici. La riflessione che proponiamo, con alcune precisazioni, può essere estesa anche alla pausazione: per quanto dall'analisi del parlato poetico rosselliano emerga un uso della pausazione decisamente strutturante e legato a fattori stilistici (il silenzio è uno dei perni del *nuovo ordine* metrico della poetessa), vi sono nelle letture molti passaggi in cui il ricorso a pause con una determinata durata risulta insolito, slegato e indipendente da precise esigenze poetiche, che invece avrebbero richiesto (o meglio suggerito) altre durate. Nelle interpretazioni dei nove lettori, il ventaglio pausale è invece più ristretto, tipico e regolare: risponde tendenzialmente a esigenze stilistiche e, inoltre, non è presente lo stesso numero di glottalizzazioni. Inoltre, è interessante osservare che nel *modus legendi* rosselliano questi fenomeni discussi e analizzati non si presentano isolatamente, ma sembrano anzi cooccorrere insieme ad altre caratteristiche note come tratti del parlato parkinsoniano, che

potrebbero per questo rafforzare l'ipotesi sulla natura patologica dell'eloquio della poetessa. Tra questi, ad es., vi sono una tendenziale ipoarticolazione e una qualità della voce che soprattutto nelle letture tarde si fa più soffiata: caratteristiche già ampiamente discusse in letteratura (Logeman *et al.*, 1978; Hanson *et al.*, 1984) e che possono essere motivate pensando alla ridotta mobilità del sistema motorio della poetessa e alle conseguenze sulla produzione dei gesti articolatori.

Circa questi aspetti, l'analisi acustica della lettura poetica di Rosselli – per quanto si tratti di un parlato pianificato – può sicuramente rilevare fenomeni di riduzione e di sottospecificazione; inoltre, non sono pochi i casi di elisione di sillabe o di coalescenza vocalica in *loci* in cui questi fenomeni non sono previsti né metricamente motivati. In controtendenza rispetto a quanto riportato in letteratura sui pazienti parkinsoniani, sono invece i dati relativi a intensità e frequenza fondamentale: come mostra il *VIP-Radar*, questi indici rivelano, infatti, un eloquio tutt'altro che appiattito e poco colorato, che anzi risulta – al contrario di quanto si potrebbe prevedere in un quadro di disartria di tipo ipocinetico – peculiarmente connotato da una grande variabilità (ampio *Pitchspan*).

7. *Conclusioni*

Attraverso l'utilizzo della metodologia di VIP, in cui conoscenze provenienti da varie discipline sono conciliate con la fonetica sperimentale, siamo stati in grado di analizzare ed estrarre parametri salienti che ci hanno permesso di ottenere una descrizione qualitativa delle caratteristiche distintive della voce e dello stile di lettura di Rosselli. Allo stesso tempo, abbiamo potuto confrontare i dati estratti in modo quantitativo, riuscendo a individuare ed esplorare le variazioni e le costanti all'interno delle interpretazioni rosselliane in relazione al tempo, alle diverse raccolte poetiche prese in considerazione e allo stile metrico dei componimenti selezionati. Tra le caratteristiche rilevate, abbiamo descritto le specifiche scelte di riproduzione degli *enjambement* adottate dalla poetessa, attraverso le quali vengono generati forti effetti di asimmetria tra il piano metrico e quello prosodico. L'ascolto delle otto interpretazioni ci ha, inoltre, permesso di individuare ridondanti e peculiari *pattern* melodico-intonativi nello stile interpretativo di Rosselli, riflessi evidenti della sua formazione musicologica e della sua attenzione ai suoni della lingua. Non meno centrale è stata la riflessione sull'interessante e caratteristico uso delle pause, dal quale è stato poi possibile prendere spunto per la seconda parte della ricerca. Non solo abbiamo mostrato come Rosselli ricorra spesso alla pausazione per generare o rafforzare passaggi poetici fonosimbolici – rendendo così più dinamiche e colorate le interpretazioni – ma abbiamo anche potuto correlare la gestione delle pause con la peculiare lunghezza del respiro dell'autrice, che soprattutto in alcune letture diventa un vero e proprio tratto distintivo e che, come abbiamo visto, potrebbe tradire una natura patologica del suo parlato. Per quanto riguarda la seconda parte della ricerca, insieme al confronto con le voci contemporanee, abbiamo mostrato tratti specifici per la prosodia di Rosselli, nettamente diversi rispetto a quelli di altri interpreti. Alla

luce di queste consapevolezze, resterebbero tuttavia da valutare le produzioni anche in considerazione di una diversa velocità d'eloquio.

Rimandando a futuri studi l'eventuale avvaloramento della natura patologica dell'eloquio rosselliano – o quantomeno la possibilità di disambiguare con più decisione tra stile e parlato patologico – con questo contributo abbiamo tentato di sollevare una riflessione su aspetti ancora inesplorati, approfondendo attraverso l'indagine fonetica il dato acustico di una delle voci più importanti e influenti del Novecento italiano. La nuova prospettiva con cui abbiamo proposto di studiare la voce di Rosselli, oltre a far emergere uno stile prosodico originale, ci ha permesso di individuare in un parlato poetico inconfondibile e atipico alcuni riflessi associabili a tratti di parlato patologico di soggetti parkinsoniani, per quanto non sia trascurabile e da escludere la concomitanza della schizofrenia. Auspicato e probabilmente risolutivo potrebbe essere un confronto tra gli indici estratti dalle letture poetiche della poetessa e da tipologie di parlato meno pianificate, come quello radiofonico o televisivo (programmi televisivi, interviste radiofoniche e docu-film in cui Amelia Rosselli dialoga e interviene). Uno studio del genere – da sviluppare con un approccio interdisciplinare, concentrando l'attenzione sulla delicata dimensione patologica del parlato e sul valore del silenzio (Albano Leoni, 2022) – potrebbe permetterci di ampliare questa ricerca tanto embrionale quanto già interessante e densa di spunti di riflessione[9]. Parallelamente a queste indagini sarebbe, infine, auspicata anche una ricerca sul parlato poetico di Rosselli in diacronia, per affinare l'analisi sugli sviluppi delle sue patologie: impressionisticamente, infatti, si possono rilevare già delle differenze tra le prime letture del corpus (del 1985) e l'ultima (del 1990). Uno studio concentrato su questi aspetti – a cui potrebbe essere affiancata anche un'analisi di tipo gestuale, da tempo tra le idee dei membri del progetto VIP – potrebbe non solo garantire una risposta più accurata ai quesiti di ricerca che questo lavoro ha aperto, ma sosterrebbe anche l'avvio di una nuova pista d'indagine sul rapporto tra intonazione, gesto e significato nel parlato poetico.

Riferimenti bibliografici

Albano Leoni, F. (2022). *Voce. Il corpo del linguaggio*. Roma: Carocci.

Bertinetto, P.M., Magno Caldognetto, E. (1993). Ritmo e intonazione. In Sobrero, A.A. (a cura di), *Introduzione all'italiano contemporaneo. Le strutture*. Roma-Bari: Laterza, 141-192.

Bertoni, A. (2006). *La poesia. Come si legge e come si scrive*. Bologna: Il Mulino.

Blankenship, J., Kay C. (1964). Hesitation phenomena in English speech: A study in distribution. *Word*, 20, 360-372.

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [Computer program] Version 6.1.41. https://www.praat.org/

[9] Ringraziamo i revisori per le precise riflessioni, che ci hanno aiutato a rivedere alcune sfumature dell'ipotesi iniziale, e per gli spunti di ricerca futuri, che auspichiamo di percorrere.

Caputo, F. (a cura di) (2004). *Una scrittura plurale. Saggi e interventi critici*. Novara: Interlinea.

Cavarero, A. (2003). *A più voci. Filosofia dell'espressione vocale*. Milano: Feltrinelli.

Ciaurelli, L. (2020). «Il linguaggio nel decadimento cognitivo: marker linguistici e automazione della diagnosi». Tesi di Dottorato. Università di Roma – La Sapienza. Tutor: Proff. F. Tamburini, A. De Dominicis, M. Falcone.

Colonna, V. (2021a). "Voices of Italian Poets". Analisi fonetica e storia della lettura della poesia italiana dagli anni Sessanta a oggi. Tesi di Dottorato inedita. Università di Genova-Università di Torino. Tutor: Prof. Antonio Romano.

Colonna, V. (2021b). La lettura della poesia italiana del secondo Novecento: una proposta di studio fonetico. In *L'analisi linguistica e letteraria, XXIX*, 5-26.

Colonna, V. (2022). *Voices of Italian Poets. Storia e analisi fonetica della lettura della poesia italiana del Novecento*. Alessandria: Edizioni Dell'Orso.

De Iacovo, V., Colonna, V. & Romano, A. (2020). La pausazione. *Bollettino del Laboratorio di Fonetica Sperimentale «Arturo Genre»*, n. 5, 41-48.

Dovetto, F.M., Gemelli, M. (a cura di) (2013). *Il parlar matto. Schizofrenia tra fenomenologia e linguistica: il corpus CIPPS*. Roma: Aracne.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, 25, 11-28.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. Tesi di Dottorato. Department of Computer and Information Science. University of Linköping, Svezia. Tutor: Proff. L. Ahrenberg e N. Dahlbäck.

Giannini, A. (2001). Corpus AVIP: ehm, ehm. In Magno Caldognetto, E., Cosi, P. (a cura di), *Multimodalità e Multimedialità nella Comunicazione*. Padova: Unipress, 179-184.

Giannini, A. (2003). Vocalizzazioni e prolungamenti vocalici. In Cosi, P., Magno Caldognetto, E. & Zamboni, A. (a cura di), *Voce, canto, parlato. Scritti in onore di Franco Ferrero*. Padova: Unipress, 163-172.

Giovannuzzi, S. (a cura di) (2012). *Rosselli. L'opera poetica*. Milano: Mondadori.

Goldman-Eisler, F. (1958). Speech Analysis and Mental Processes. *Language and Speech*, 1(1), 59-75.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Goldman-Eisler, F. (1972). Pauses, Clauses, Sentences. *Language and Speech*, 15, 103-113.

Hanson, D.G., Gerratt, B.R. & Ward, P.H. (1984). Cinegraphic observations of laryngeal function in Parkinson's disease. *Laryngoscope*, 94(3), 348-353.

Heeman, P.A., Allen, J.F. (1999). Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialog. *Comp. Linguistics*, 25(4), 527-571.

Kendall, T. (2009). Speech rate, pause and linguistic variation: an examination through the sociolinguistic archive and analysis project. Tesi di Dottorato, Department of English. Duke University, Durham, USA. Tutor: Prof. Walt Wolfram.

Lickley, R.J. (1994). Detecting Disfluency in Spontaneous Speech. Tesi di Dottorato. Dep. of Linguistics & Centre for Speech Technology Research. University of Edinburgh, UK. Tutor: Prof: E. Bard.

Logemann, J.A., Fisher, H.B., Boshes, B. & Blonsky, E.R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *J Speech Hear Disord*. 43(1), 47-57.

Lo Iacono, F. (2023). «La musica comunque fa la sua parte»: analisi fonetica dello spazio metrico rosselliano attraverso la metodologia di *Voices of Italian Poets*. Tesi di Laurea Magistrale inedita. Università di Bologna.

Mahl, G.F. (1956). Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology*, 53(1), 1-15.

Manera, M. (2006). «Le più fantastiche imprese: spazi versi rime tempi». Il sistema metrico di Amelia Rosselli. Tesi di dottorato, Università degli studi di Torino.

Menichetti, A. (1993). *Metrica italiana. Fondamenti metrici, prosodia, rima*. Padova: Antenore.

Nodari, R., Calamai, S. (2021). I silenzi dei matti. Gli spazi 'vuoti' del parlato nell'archivio sonoro di Anna Maria Bruzzone. *Abstract dell'intervento al XVII Convegno AISV (Zurigo, 04-05/02/2021)*.

O'Shaughnessy, D. (1992). Analysis of false starts in spontaneous speech. *Proceedings of the 2nd International conference on spoken language processing*. Banff, Canada, 2-5.

Paris, R. (2020). *Miss Rosselli*. Vicenza: Neri Pozza.

Romano, A. (2008). *Inventari sonori delle lingue: elementi descrittivi di sistemi e processi di variazione segmentali e sovrasegmentali*. Torino: Edizioni dell'Orso.

Rosselli, A. (1962). Spazi metrici. In Id. *Le poesie*. Milano: Garzanti, 1997, 337-342.

Rosselli, A. (1977). Storia di una malattia. In *Nuovi Argomenti*, 56: 185-196.

Savy, R. (2006). Specifiche per la trascrizione ortografica annotata dei testi raccolti. In Albano Leoni, F., Giordano, R. (a cura di), *Italiano parlato. Analisi di un dialogo*. Napoli: Liguori, 1-37.

Savy, R. et al. (2006). Specifiche per la trascrizione e l'etichettatura dei livelli segmentali in CLIPS. In *CLIPS – Corpora e Lessici di Italiano Parlato e Scritto* (www.clips.unina.it).

Sessa, P. (2018). *La lettura, il corpo, la voce. Fondamenti linguistici e neurali della lettura ad alta voce*. Roma: Giovanni Fioriti Editore.

Shriberg, E.E. (1994). Preliminaries to a theory of speech disfluencies. Tesi di Dottorato. University of California, Berkeley, USA. Tutor: Prof. S. Ervin-Tripp.

Venturini, M., De March, S. (2010). *Amelia Rosselli. È vostra la vita che ho perso. Conversazioni e interviste 1964-1995*. Firenze: Le Lettere.

DOI: 10.17469/O2111AISV000010

CLAUDIO ZMARICH, SIMONA BERNARDINI, SABRINA BONICHINI,
MIRKO BONOTTO, FEDERICA CHIARI, ERIKA GANASSIN,
CATERINA PISCIOTTA, FABIOLA VULLO, ELISA ZACHEO

Linguistic proficiency, communicative attitude, and severity in a sample of children who stutter aged 4-6 years

Children who stutter develop self-awareness of their speech from an early age, resulting in a more negative communication attitude (CA) compared to their fluent peers. Other speech disorders can also affect CA. The relationship between low scores in linguistic proficiency and negative CA among children who stutter without other comorbidities is still unclear. Eighteen children (aged 4;0-6;11) diagnosed with stuttering, but not yet treated, Italian speakers from Veneto, with no other concurrent disorders were assessed using KiddyCAT for CA and BVL_4-12 for linguistic proficiency. Their linguistic abilities were comparable to those of fluent peers, except for their poor performance on articulation and repetition tests. Although most children stutter moderately or severely, their CA was comparable to that of their normal, fluent peers. No significant correlations emerged between CA, language ability, and stuttering severity.

Keywords: children who stutter, Communicative Attitude, stuttering severity, linguistic proficiency, non-words repetition.

1. *Introduction*

Although the International Classification of Diseases has currently arrived at the 11th edition (ICD-WHO, 2021), we still prefer the definition of stuttering contained in the 9th: "disorders in the rhythm of speech, in which the individual knows precisely what he wishes to say, but at the time is unable to say it because of an involuntary, repetitive prolongation or cessation of a sound" (ICD-WHO, 1977, p. 202). Our preference is motivated by the central role assumed here by the "involuntary" character of the stuttering's disfluencies. One can easily imagine what kind of despair and frustration can be induced by such an experience of involuntary loss of control (Perkins, 1990). These negative feelings take place from the beginnings of stuttering for most of the children who stutter (CWS), and then they give rise to a constellation of reactions which are an integral part of the stuttering syndrome and have received the label of "secondary", or "associates", symptoms (Yairi, Seery, 2023).

Secondary symptoms are primarily caused by a negative communication attitude (CA) towards one's own speech, along with an image of oneself as a poor/bad speaker. In the upcoming exposition, we will address: 1) the concept of CA and its significance for CWS; 2) its applicability to other speech and language disorders,

not just stuttering; 3) the high rate of comorbidity of CWS with speech and language disorders; 4) the speech and language factors interacting with stuttering in early childhood.

1.1 Communicative Attitude and stuttering

The close relationship between stuttering and CA has been the subject of important research and debates in recent decades. To date, clinicians have embraced the idea that stuttering should be considered as a complex and composite disorder. The manifest symptoms, in fact, represent the central, but not the only, attribute of this multifaceted pathology, which also includes the experience of the person who stutters (PWS). According to this perspective, the evaluation and treatment of PWS must be addressed to the disfluency and to the negative emotional reactions, to the secondary dysfunctional coping behaviours attempted by the subject to overcome and/or avoid stuttering, and to the cognitive dimension. The therapeutic process can only be considered positive and satisfactory if verbal disfluency is reduced and the CA is normalized. (Menzies, O'Brian, Onslow & Packman, 2008). In fact, negative CA appears to play a fundamental role in the maintenance of the disorder (Guttormsen, Kefalianos & Næss, 2015; Bloodstein, Bernstein-Ratner & Brundage, 2021).

Research has established that stuttering typically starts in preschool years (Yairi, Seery, 2023) and that children could be aware of the dysfluency about their speech as early as they are three years old, at the same time or shortly after the stuttering onset, and this awareness increases along with age. Hence, CWS not only can be aware of their disfluency from the beginning, but they also develop a perception of their disfluency as negative. This negative speech-related attitude includes viewing themselves as bad communicators, evaluating speech as difficult, being apprehensive in speaking situations and feeling high anxiety when communicating with others. (Ambrose, Yairi, 1994; Blood, Blood, Tellis & Gabel, 2001; Ezrati-Vinacour, Platzky & Yairi, 2001; Mulcahy, Hennessey, Beilby & Byrnes, 2008; Iverach, Rapee, 2014). This can lead CWS to implement waiving and communication delegations to alleviate the anxiety due to the fear of being judged negatively by the interlocutor. Unfortunately, a communication disability such as that caused by stuttering, can have negative repercussions on the overall quality of life of the PWS throughout the lifespan (Yaruss, Quesal, 2004). These include limitations in social interactions, prejudices and social stigma, distress, shame, and low self-esteem (Cream, Onslow, Packman & Llewellyng, 2003; Ginsberg, Wexler 2000). Stuttering sometimes leads PWS to encounter limitations in school career and career choices (Daniels, Gabel & Hughes, 2012; Gabel, Blood, Tellis & Althouse, 2004). Not only CWS but also the children who do not stutter (CWNS) can be aware, as early as when they are three years old, of the stuttering-like-disfluencies (SLD) produced by CWS, and they can give them a negative judgement. As a matter of fact, CA tends to be more negative in CWS compared to CWNS (Guttormsen *et al.*, 2015).

A measure of communication attitudes, such as the *Communication Attitude Test for Preschool and Kindergarten Children Who Stutter* (KiddyCAT, Vanryckeghem, Brutten, 2007), has been shown to meet the criteria of assessing the attitudes towards communication in preschool-age children. Recently, KiddyCAT has been translated and adapted to Italian (Vanryckeghem, Brutten, 2022).

It had been speculated that factors such as age, severity, history of stuttering (e.g., time since onset), gender, parental influences, metalinguistic awareness, treatment and language proficiency may influence awareness of stuttering and negative attitude about speech (Ambrose, Yairi, 1994; Yairi, Seery, 2023; Boey, Van de Heyning, Wuyts, Heylen, Stoop & De Bodt, 2009).

There are contradictory results regarding age. A number of studies found that age did not significantly affect the difference in CA between CWS and CWNS (Vanryckeghem *et al.*, 2005; Bernardini *et al.*, 2009; Bernardini *et al.*, 2019). However, other studies have shown that the CAT scores of CWS tend to increase with age (Kawai *et al.*, 2012), while those of CWNS tend to decrease (Vanryckeghem & Brutten, 1997; Brutten, Vanryckeghem, 2007; Clark, Conture, Frankel & Walden, 2012), leading to the result that the differences in CA between CWS and CWNS increase with age. The review by Guttormsen *et al.* (2015) supports the hypothesis that the CA of CWS worsens with increasing age. The authors suggest that CWS may be exposed to increasingly negative communication experiences as they grow older: for example, they may experience difficulties in school, where communication needs become increasingly demanding, and where they are also at greater risk of being bullied (Blood, Blood, 2004; Cook, Howell, 2014). However, not all studies conclude that CA gets worse with age. For example, in the study by Clark *et al.* (2012) the group of younger children (from 3;0 to 4;6 years) obtained higher scores (more negative attitude) on the KiddyCAT than that of the older ones (from 4;7 to 5;11 years), and Groner *et al.* (2016) also found that CA improved with increasing age in both CWS and CWNS. The age where CWS scores start to differ from those of CWNS has been investigated in research on CA during the developmental period. Thanks to the use of the KiddyCAT, it was found that the CA of the CWS and that of the CWNS begin to differ around the age of 3-4 years, when stuttering usually begins (Bernardini *et al.*, 2009).

As to severity, in adults and adolescents, this is not linearly correlated with either CA or the overall impact of the disorder. Specifically, for some PWS the severity is mild, but the impact of the disorder is high, while for others the opposite may occur (Winters, Byrd, 2021). Conversely, research in school-aged children suggests that there may be a relationship between CA and severity, although there are some conflicting findings. Ambrose, Yairi (1994) did not find any relationship between the severity of stuttering and stuttering speech awareness. More recent studies suggest the existence of such a relationship, as they found that children with more severe overt symptoms had more negative CA (Beilby, Byrnes & Yaruss, 2012; Kawai, Healey, Nagasawa & Vanryckeghem, 2012; Vanryckeghem *et al.*, 2001). However, other studies do not confirm this relationship: Groner *et al.* (2016)

provided evidence both in favour and against it, while in the study by Winters, Byrd (2021), none of the measured characteristics of severity (frequency and duration of disfluencies, physical concomitants) correlated with CA. Given these conflicting results, we cannot conclude that CA is associated with stuttering severity, but neither can we rule out this possibility. If this hypothetical relationship could be confirmed, it would help clinicians to identify early children who need immediate treatment (Guttormsen *et al.*, 2015).

Regarding the relationship between CA and gender, there is consensus on the fact that CAT or KiddyCAT scores do not differ significantly between boys and girls who stutter (Vanryckeghem *et al.*, 2005; Bernardini *et al.*, 2009; Clark *et al.*, 2012), as confirmed by a recent meta-analysis (Guttormsen *et al.*, 2015).

1.2 Communicative Attitude and Speech and Language disorders different from stuttering

Despite the wide debate on the subject, numerous studies support the hypothesis that CWS have a high probability of exhibiting concomitant disorders, especially of a linguistic nature (Blood, Ridenour, Qualls & Hammer, 2003). In the light of the fundamental role played by CA in stuttering, it seems useful to review some of the main evidence in the literature regarding the characteristics assumed by CA in other speech and language disorders.

The study by De Nil, Brutten (1990), using the CAT on children aged 7 to 14 years, allowed to compare the CA of the CWS with that of peers affected by other speech disorders. Participants were divided into four groups based on the disorder they experienced, including stuttering, articulation problems, voice problems, and a control group. The results highlighted the presence of more negative attitudes in all speech disordered groups than in the control group, although the highest scores were obtained by CWS and by children affected by voice disorders (De Nil, Brutten, 1991). Evidence of negative CA has also been reported in 10-year-old children with cleft lip and palate (Havstam, Sandberg & Lohmander, 2011). They scored higher than their normal fluent peers on the Swedish version of the CAT test (CAT-S) (Johannisson, Wennerfeldt, Havstam, Naeslund, Jacobson & Lohmander, 2009).

A study analyzed the CA of a group of children with speech sound disorders (SSD) in school and preschool age using the KiddyCAT test and *The Child Form of the Speech Participation and Activity Assessment for Children* (SPAA – C) (McLeod, 2004; cited in McCormack *et al.*, 2019), an informal questionnaire about how the child perceives his own speech. The results revealed that 58.9 % of the participants achieved a score on the KiddyCAT index of positive CA, while the remaining 41.1 % achieved a score much higher than expected for the corresponding age, thus showing a tendentially negative attitude. Furthermore, by relating the KiddyCAT data to the child's articulatory accuracy, it emerged that the more negative CAs were correlated with lower phono-articulatory skills, expressed by a lower *Percentage of Correct Consonants* (PCC) (McCormack *et al.*, 2019). The study then examined the relationship between the parents' assessment of the child's intelligibility, the

perception they reported of their children's CA, and the attitude that the latter showed in the KiddyCAT. The results indicated a substantial discrepancy between these measures, revealing that the child's CA was not correlated with that perceived by the parent nor with the level of intelligibility reported by them. This represents a further confirmation of how, in clinical practice, it is necessary to adequately take into consideration the point of view of the child, as well as that of the parent (McCormack *et al.*, 2019).

The possibility that disfluency awareness and negative CA could not to be associated in a direct way is suggested in a study by Choi, Hurd & Kim (2022) on CWNS: they found that those with higher language ability are more likely to be aware of, and negatively evaluate, disfluent speech than the children with lower language ability. Perhaps, children with higher, compared to lower, language abilities seem more aware of differences and deviations in speech disfluencies, because they have a better speech language monitoring system to monitor their speech output in terms of phonology, morphology syntax, semantics, and fluency, so as to produce error-free speech (Levelt, 1993). This result suggests that children affected by speech and/or language disorders (like stuttering) may not have a speech-language monitoring system sufficiently developed, which may cause them not to be aware of their problems and consequently not to be negatively affected in their CA.

1.3 Comorbidities of stuttering with speech and language disorders

Current research highlights the presence of concomitant speech and language disorders and/or other disorders in approximately two-thirds of CWS (Blood *et al.*, 2003). However, this figure is affected by numerous differences between the various studies, attributable to the different methodologies used and the composition of the samples, in terms of age of the participants and diagnostic inclusion criteria (Arndt, Healey, 2001).

From a clinical point of view, studies concerning comorbidity with stuttering assume considerable importance, since the evaluation and treatment of children with concomitant disorders has very different connotations compared to those relating to isolated stuttering (Blood *et al.*, 2003).

Furthermore, according to some authors, phonological disorders associated with stuttering in preschool age are a potential negative prognostic indicator of stuttering recovery (Conture, 2001), although there is evidence to the contrary (Watkins, Yairi, & Ambrose, 1999).

Historically, the hypothesis of comorbidity between speech disorders and stuttering refers to the DC Model (Demands and Capacities Model; Adams, 1990; Starkweather, Gottwald, 1990), which argues that the conversation demands can induce an overload of the language production systems of the child and a consequent increase in disfluent episodes when they exceed his motor, linguistic and emotional features. Arndt and Healey (2001) highlighted the presence of associated linguistic disorders in 44 % of 205 CWS aged between 3 and 20 years, with a share of phonological disorders equal to 33 %. Data seems to support the

hypothesis of a higher incidence of phonological disorders in CWS compared to the general childhood population (20 % of associated phonological disorders in CWS compared to 2.6 % among CWNS) (Arndt, Healey, 2001; Blood *et al.*, 2003).

Blood *et al.* (2003) reported the presence of associated linguistic disorders in 62.8 % out of a sample of 2628 CWS, with a percentage of articulation disorders equal to 33 % and 12.7 % of phonological disorders. The latter did not correlate with the level of stuttering severity (Walk, Edwards & Conture, 1993; Yaruss, Conture, 1996). Other linguistic disorders frequently found in CWS concern the lexical-semantic production (13.5 %) and lexical-semantic understanding (12.1 %) (Blood, *et al.*, 2003), as well as voice problems (0.12 %) and SSD (Speech Sounds Disorders), with a prevalence of 1.06 % (McKinnon, McLeod & Reilly, 2007).

In contrast to the research mentioned above, the review by Nippold (2012) and a study by Smith and Weber (2017) suggest that CWS, as a group, are no more prone to language impairment than CWNS, but report heterogeneous linguistic abilities placing themselves on all levels of performance.

1.3.1 Comorbidities of stuttering with disorders different from speech and language disorders

According to Blood *et al.* (2003), 34.3 % of the sample shows concomitant disorders with primary causes outside from the speech and language ability. Specifically, they involve: 15.2 % of cases with associated DSA (8.2 % of which diagnosed with dyslexia and/or dysorthography), and 5.9 % with concomitant ADHD. Among the co-occurring non-language disorders, five were found to have a frequency of less than 5 %. They include: central auditory processing disorders (3.1 %), neuropsychological disorders (2.9 %), behavioural disorders (2.4 %), sensory integration disorders (2.1 %) and acquired disorders (1.3 %). The incidence of concomitant disorders (both linguistic and non-linguistic) affects males more than females, especially for concomitant speech-sound disorders (48.2 % of males against 38.8 % of females) and learning disorders (12.2 % of boys versus 8.3 % of girls) (Blood *et al.*, 2003).

The systematic analysis of health information by the electronic health record (*EHR*) provided a more comprehensive view of the comorbidity between stuttering and other disorders: it compares 572 subjects diagnosed with developmental stuttering with 2765 normal fluent subjects. The aim of the study was to verify the possibility of identifying possible disease associations existing with stuttering, using a study approach known as *PheWAS* (Phenome-Wide Association Study), that uses *Phecode* (labels inspired by the ICD 9 classification), including all the phenotypic information with common characteristics. In this regard, associations were found with Neurodevelopmental Disorders and Language Disorders, followed by Pervasive Developmental Disorders and ICT (Pruett, Shaw, Chen, Petty, Polikowsky, Kraft, Jones & Below, 2021). Autism spectrum disorder does not appear to be in statistically significant comorbidity with stuttering, while the strong co-presence of deafness and stuttering is in contrast with most of the evidence in the literature, reporting a low prevalence of hearing loss in the stuttering population (Montgomery, Fitch, 1988; Yairi, Seery, 2015).

1.4 Speech and language factors impacting stuttering in childhood

There are many ways to evaluate speech and language proficiency in CWS with respect to CWNS. The first distinction to be made is between behavioural and imaging measures. Due to space restrictions, we will give prominence to the behavioural studies. The reader interested in brain imaging and electrophysiological studies can refer to Chang *et al.* (2018) and Chow *et al.* (2023), among others. Within the behavioural studies, we can further distinguish among studies applying standardised/commercial tests (or batteries) to samples of CWS and CWNS (community-screening studies included), and studies using nonstandard experimental methodologies to assess one or more specific capacities of CWS.

1.4.1 Standardised test batteries

The study by Coulter, Anderson & Conture (2009), conducted using various standardised tests for the evaluation of the linguistic abilities of CWS and CWNS, showed the CWS group to be higher than the norm as to general linguistic abilities, but lower than the norm as to those more frankly phonetic-phonological. The research followed the model of the study by Anderson, Pellowski & Conture, (2005), according to which the CWS have, on average, three times the probability of showing dissociations between the different language skills compared to CWNS. A meta-analysis of 22 studies investigating the language skills of CWS and CWNS aged 2-8 years also revealed subtle differences between these two groups in measures of general language (MLU), and receptive and expressive vocabulary. The performance of the CWS, although within the norm values, was in fact poorer on average than that exhibited by CWNS (Ntourou, Conture & Lipsey, 2011). It has been hypothesised that the frequent dissociations between the single linguistic levels found in CWS could make their linguistic processing system more vulnerable to interference on the cognitive, sensory-motor and emotional level during conversation, increasing the likelihood of experiencing disfluent episodes (Bosshardt, 2006; Coulter, Anderson & Conture, 2009; Clark, Conture, Walden & Lambert, 2015; Smith, Weber, 2017).

Numerous studies have placed CWS into different subgroups based on the language skills they exhibit (Yairi, 2007; Hubbard-Seery, Watkins, Mangelsdorf & Shigeto, 2007; Reilly *et al.*, 2009). A first subgroup includes CWS with poor language skills and high comorbidity for language disorders (Conture, Zackheim, Anderson & Pellowski, 2004; Hollister *et al.*, 2017; Walden *et al.*, 2012), while a second subgroup includes CWS who obtain better scores than CWNS in the main standardised tests for the assessment of linguistic skills and in all measures of expressive language (Watkins, Yairi & Ambrose, 2005; Watts *et al.*, 2017).

These conclusions have been criticised by Nippold (2012; 2019), as many of the studies included by Ntourou, Conture & Lipsey (2011) were not adequately balanced in the sample recruitment criteria, the socioeconomic characteristics of the participants (SES) and the different language tests administered. Nippold's meta-analysis reported opposite conclusions to those of Ntorou and collaborators,

highlighting how CWS show linguistic skills equal or superior to those exhibited by CWNS of the same age, although these cannot be defined as exempted from linguistic weaknesses (Nippold, 2012; 2019). Indeed, there is evidence demonstrating the presence of impaired language skills in some CWSs (Smith *et al.*, 2012), like what can be found in some CWNSs (Leonard, 2014). Perhaps the most balanced and convenient description has been put forward by Smith, Weber (2017): CWS as a group would obtain highly heterogeneous performances, with higher percentages of individuals ranking at an extremely good or extremely poor level (Yaruss, LaSalle & Conture, 1998; Hubbard-Seery *et al.*, 2007).

Despite the lack of evidence suggesting that CWS are truly language-disordered, numerous studies have demonstrated subclinical impairment of language skills in cohorts of CWS. Two relatively recent community-screening studies (Zaretsky *et al.*, 2017; Shimada *et al.*, 2018), were able to compare language skills in CWS with much larger cohorts of CWNS in the same jurisdiction, all of whom were screened as part of public health projects. Both found relative depression of oral language skill in CWS, especially regarding articulation and grammar.

In contrast, an Australian community health project, the Early Language in Victoria Study (ELVS), using different screening methods, found a slight oral language advantage, rather than disadvantage, for CWS that they tracked, compared with children who were never reported as stutterers (Kefalianos *et al.*, 2017). However, one large-scale study of 100 pairs of CWS and CWNS collected across numerous American laboratories (Luckman *et al.*, 2020) found that CWS scored within 1 SD lower in both expressive and receptive vocabulary, as measured by standardised assessments, than did appropriately matched fluent peers.

1.4.2 Phonetic and phonological levels

According to Sasisekaran (2014), it is possible to distinguish at least three different points of view on the role of Phonology in stuttering (she does not consider the level of articulatory planning and delivery):

1. the study of the phonological structures that predict the placement of SLDs in the speech chain;
2. the exploration of phonological competence through standardised tests;
3. the study of cognitive processes related to phonological processing (retrieval and construction of a phonetic plan from the lemmas to the utterance);

As to 1), in the words of Brundage, Bernstein-Ratner (2022: 8), "adult speakers are more likely to stutter on words beginning with consonants, on longer words, on content words, and on words occurring early in the utterance. Long and complex utterances also increase stuttering frequency when compared with short, simple utterances", and CWS differ from adults only for the tendency to stutter on function words (Howell, Au-Yeung, 1995; see also Bloodstein, Ratner & Brundage, 2021). The literature has often explored the link between the articulatory complexity of the word to be produced and the likelihood of encountering disfluent episodes., noting for instance how CWS usually make less use of complex consonants than CWNS (Howell, Au-Yeug, 1995) and how SLDs occur more frequently in

word-initial position, and more in stressed than in unstressed syllables (Howell, Au-Yeung & Sackin, 2000; Natke, Sandrieser, Van Ark, Pietrowsky & Kalveram, 2004). Various studies also report the presence of subtle differences in motor aspects between the speech of CWS and that of CWNS. These differences include laryngeal function (reduced amplitude and intonation range in phonation), a lower amplitude of articulatory movements and a lower speed of speech (Yairi, Ambrose, 2005). The authors support the hypothesis of poor motor coordination stability and less flexibility in speech planning in preschool CWS: the speech of these subjects appears in fact more vulnerable to variations in the speech rate, linguistic complexity, and social and cultural pressure (Smith, Zelaznik, 2004; Walsh, Mettel & Smith, 2015).

These data have broad clinical utility, as they allow early prediction of the probability of remission or persistence of the disorder. According to literature, CWS who have better articulatory skills are more likely to experience spontaneous recovery than those who have lower scores on linguistic articulation tests (Spencer, Weber-Fox, 2014). Conversely, CWS presenting an excessive degree of intrasyllabic coarticulation for a period longer than the 6th/7th month from the onset, would more frequently encounter the persistence of the disorder (Subramanian, Yairi & Amir, 2003).

The twenty-year-long research project by Yairi and colleagues has also shown that CWS, two years after the onset, are on average later than CWNS in phonological development (Paden, 1996). Notably, some authors have reported a more diverse typology of phonological processes in the speech production of preschool CWS than in CWNS. These processes are often age-inappropriate or atypical, although they are rarely categorised as a real disorder (Louko, Edwards & Conture, 1990).

The hypothesis of a greater immaturity in the phonological processing of preschool CWS was instead supported by the research of Byrd, Conture & Ohde (2007), related to the facilitating effect (priming) of lexical access following holistic acoustic stimuli (all the disyllabic word minus the initial consonant) and incremental (only the initial consonant and the transition to the next vowel). These studies have shown how the first type of elaboration, typical of early developmental stages, facilitates the recovery of the lexical target in children with and without stuttering already at the age of 3, while the second, typical of the adult population, appears to be only advantageous for CWNS ages 5 and up.

According to psycholinguistic theory, the disfluencies would also be caused by a phonological coding deficit, related to the recovery and construction of the phonological segments of words (Pelczarski, Yaruss, 2016). A fundamental role in this regard is played by the phonological memory, which allows the sound representation to be formed and preserved for a sufficiently long time to be correctly reproduced (Pelczarski, Yaruss, 2016). It is usually tested by non-word repetition trials, in which CWS perform on average worse than CWNS (Hakim, Bernstein-Ratner, 2004; Anderson *et al.*, 2006; Anderson, Wagovich & Brown, 2019). The difference assumes significance especially for the CWS under the age

of 4-5 years with a greater risk of becoming persistent (Bankson, Bernthal, 1990), and is detected more frequently on trisyllabic words, like what is evidenced in the population with a diagnosis of DPL (formerly DSL) (Hakim, Bernstein-Ratner, 2004). Furthermore, as with DPL, the phonological memory impairment found in CWS may have cascading effects on lexical and syntactic development, limiting the acquisition and retention of new words and their subsequent organisation into sentences (Coady, Evans, 2008).

1.4.3 Syntactic and morphological levels

The influence exerted by morphological and syntactic variables on the probability of incurring disfluent episodes is now known in the literature. In particular, in the CWS population, disfluencies have been shown to occur more frequently in sentence-initiating words (Richels *et al.*, 2010; Buhr, Zebrowski, 2009), and in function words rather than content words (Howell, Au-Yeung, & Sackin 2000). Furthermore, the onset of the disorder occurs when the acquisition of the syntactic ability to produce the utterance is developing, between the ages of 2 and 4 years. In this stage, the planning of the utterance switches from a purely semantic level to a syntactic one, with a discontinuity that can be translated into an increase in the frequency of disfluencies and self-corrections (Wijnen, 1990). The latter, according to some authors, could be associated with the excessive demand exerted by the growing syntactic complexity on a still underdeveloped linguistic system (Zackheim, Conture, 2003).

One of the most frequently used measures for the indirect evaluation of the child's syntactic development is the Mean Length of the Utterance (MLU), typically calculated, in Italian, as the ratio between the total number of words produced and the total number of utterances (Brown, 1973). This tool becomes unreliable in capturing the internal syntactic complexity of the utterance produced by the child when he acquires the ability to combine 3 elements in a sentence (approximately 3 years of age, Pinton, Lena, 2015).

According to the aforementioned meta-analytic review carried out by Ntourou and colleagues (2011), the spontaneous speech of CWS would see the production of on average shorter utterances (poor MLU), compared to those of CWNS. This difference may be attributable to the possible strategies for avoiding disfluencies implemented by CWS, or to their temperamental aspects (among the characteristic factors responsible for the reduction of verbal output, there would be, for example, the scarce propensity to report in an evaluation context).

1.4.4 Lexical-semantic level

Lexical skills allow to learn and distinguish the single words of a vocabulary, while semantic skills link these words to related representational and operational concepts. In the child's linguistic development, the lexical and semantic aspects are strongly connected, as initially the subject tends to produce words with an excessively generalised meaning (holophrases) and enriches his vocabulary only after repeated experiences of the world around him (Deodatus, 2014). To accurately assess these

skills, it is important to take into account the personal and social variables that affect the child's language exposure and experiences. (Luckman, Wagovich, Weber, Brown, Chang, Halle & Bernstein Ratner, 2020). It is in fact known that lexical skills tend to increase with age, are more developed in the female population at an early age (Eriksson, 2012; Luckman *et al.*, 2020), and are positively influenced by a high familiar SES (Hoff, Tian, 2005).

With regard to stuttering, many studies have reported heterogeneous results regarding the lexical skills of CWS compared to those of CWNS. Such different results could be attributable to the methodological differences found in the various studies differing on the abovementioned variables (age, gender, SES).

Among the studies that have found a poorer performance in CWS compared to CWNS there is that of Silverman, Bernstein-Ratner (2002), aimed at assessing lexical production and comprehension using standard tests. This study, conducted on a sample of 15 CWS and 15 CWNS with an average age of 35 months, reported a deficit in the development of expressive vocabulary in CWS. According to the authors, this result could reflect a strategy of avoiding the words that elicit more disfluencies and would not necessarily represent a difficulty in finding the correct lexical target by the CWS (see also Bloodstein, Ratner & Brundage, 2021). In this regard, a study by Quartarone, Bonichini, Natarelli, Pisciotta & Zmarich (2016) suggests that delays in lexical development may be present in children who subsequently exhibit stuttering well before the onset of stuttering, and children whose stuttering becomes persistent are significantly more delayed and exhibit more SLDs than those whose stuttering will recover spontaneously. Fourteen children, all born in a family where one or both relatives were PWS, were audio and video recorded at 24 months while performing an informal interview and a phonetic test. Adding to this, the MacArthur-Bates CDI was compiled by the parents. To determine the final clinical outcome of each child, a structured telephone interview was conducted with the parents after an average of 40 months from the stuttering onset (range: 20-68 months), revealing that 4 children had become persistent stutterers (CWS), and 10 had recovered spontaneously (CWNS). By comparing the lexical age of the subjects against age-referenced normative values when they were 24 months old, both groups of children presented a delay in lexical age. Still, future CWS were significantly more delayed (M = -3.42 months) than the future CWNS (M = -1.90 months), also producing a significantly higher number of SLD (3.01 % vs 0.75 %).

The studies by Anderson, Conture (2000) and by Ntourou *et al.* (2011), also reveal greater weaknesses of CWS than CWNS on all measures of receptive and expressive vocabulary. The differences between the two groups are probably attributable to the poor phonological working memory exhibited by the CWS (Gathercole, 2006; Gathercole *et al.*, 1999) and to the limitation of attention capacities, frequently found in these subjects. Attention, in particular, is considered an indispensable factor for semantic-lexical development, as it allows children to

grasp the correlation between each new word and its semantic referent (Spaulding, Plante & Vance, 2008; Finneran, Francis & Leonard, 2009).

Regarding the characteristics of the CWS group alone, the study by Coulter *et al.* (2009) highlights significant dissociations between the receptive and expressive vocabulary, demonstrating how the former is on average less developed than the latter in subjects with stuttering. This evidence, although in line with what previously reported by Anderson, Conture (2000), constitutes a unique and unexpected result, if we consider that, in the evolutionary stages of language development, verbal comprehension precedes and influences production. Interesting studies on the priming effect were then carried out to analyse the temporal course of linguistic processing, from the stage of perception of the stimulus to its articulation, through the retrieval of the lemma and the subsequent coding of the corresponding lexeme (Levelt, 1989; Bock, Levelt, 1994). A first experiment showed that CWNS usually name the target more quickly when preceded by a semantically related prime, contrary to CWS who do not seem to benefit from it and who are instead more facilitated by syntactic priming (Pellowski, Conture, 2005). This result seems to confirm the hypothesis put forward by Hall (2004) according to which important dissociations between syntactic and semantic-lexical skills in the child could affect the structuring of the utterance, interrupting or slowing down language planning and consequently leading to an increase in disfluent episodes.

Finally, the results of the aforementioned research and those of six other experimental studies were compared in the review by Luckman *et al.* (2020), which highlights that the performance of CWS between the ages of 2 and 8 is on average one standard deviation below the CWNS average value in expressive vocabulary, although it never falls into the deficit range. Furthermore, the lexical variety exhibited by CWS, measured in a conversational context, was found to be comparable to the group of CWNS. These data suggest that, during spontaneous interaction, CWS tend to select the words that best suit their coding skills, and instead find themselves in difficulty in standard tests, when they are asked to retrieve a specific lexical target.

1.4.5 Discourse level

The assessment of the child's narrative skills is of fundamental importance in the clinical setting as it investigates both the micro-processing aspects of expressive language and the macro-processing ones (Nippold, 2010; Nippold, 2012). Furthermore, narration can highlight more nuanced linguistic problems not directly identifiable by common standard tests (Roth, Spekman, 1986; Westby, 1979; Nippold, 2012).

In an initial study of the narrative abilities of school aged male CWS, Nippold, Schwarz & Jescheniak (1992) found no significant differences between them and CWNS peers in the dimensions of narrative length and complexity, of the structural integrity of the reported episodes and of the grammatical components of the story. Similar conclusions were reported by Scott, Healey and Norris (1995) on

school-age CWS who were given a retelling task, and more recently by Bajaj (2007), focused on the analysis of the narration ability on an iconic stimulus (cf. *Frog Story*).

An interesting study by Weiss, Zebrovsky (1994), conducted on 16 preschool CWS and CWNS, tested the narrative skills of the sample in the retelling task of 2 stories to listeners who already knew them or not, and in the telling of 3 invented stories. The study highlighted how, contrary to CWNS, CWS are not used to resorting to longer or richer utterances in reporting the story to those who do not know it, although they still manage to transmit the main basic information. This result suggests, on the one hand, the presence of a greater narrative effectiveness in CWS, and on the other reveals a certain pragmatic difficulty related to the inferential sphere and the amount of information to convey to each interlocutor.

1.5 Experimental hypotheses and aims

The present study aims at investigating possible correlations among stuttering severity, CA, and linguistic skills, in a group of CWS. First of all, it is assumed that CWS could develop a negative CA not only because of fluency problems, but as a result of other speech and language difficulties different from stuttering they might have, although not necessarily of clinical relevance, like the difficulties evidenced by scoring low at language batteries. Second, since we expect that negative CA increases together with the number of stressful events associated with stuttering, we expect that it could increase along with severity, age at the time of the recording, age at stuttering onset, and the amount of time elapsed from the stuttering onset.

2. *Method*

The study involved 18 children (15 males, 3 females) aged between 4 years, zero months, and 6 years, 11 months (mean age 5 years, 5 months), diagnosed as stutterers by certified speech therapists, but not yet treated. The onset of stuttering, as reported by their parents, happened at a mean age of 43 months (SD 14.3). They were Italian monolinguals, living in Veneto, without other certified concomitant disorders. The *SSI-4* (Riley, 2009) was used to assess stuttering severity, the *BVL 4-12* (a psycholinguistic battery with wide coverage, Marini *et al.*, 2015), for language and speech skills assessment in production and perception starting from four years of age, and the *KiddyCAT* (Vanryckeghem & Brutten, 2022, which is the Italian adaptation of Vanryckeghem & Brutten, 2007), for the evaluation of the CA from 3 to 6 years of age. In addition to signing the consent to participation and to data treatment, parents completed a questionnaire about the subject's medical history and family information.

The tests were administered to each child through two meetings lasting about 2 hours each, at children's home or at the speech therapist's office. All sessions were audio- and video-recorded with professional instruments. The video recordings were carried out using the Sony HDR-CX115E TV camera, while audio recordings

were acquired using the the Edirol-R09 HR, placed approximately 20-30 cm away from the child's mouth, sampling the acoustic signal at 44kHz and 16 bits.

We used the normative samples contained in the KiddyCAT (Vanryckeghem *et al.*, 2022) and in the BVL to compare the experimental subjects with a valid sample of CWNS. For this reason, the scores are expressed as z-scores.

As to severity level, there were 3 severe, 9 moderate, 4 mild and 2 very mild subjects.

3. *Results*

The average score of subjects on the KiddyCAT is 2.82 (DS=2.18; CWS' mean in Vanryckeghem, Brutten, 2022: 4.37; SD: 2.56) and it is not significantly different from the mean score of CWNS (0.72; SD: 1.19). Only 4 children out of 18 have obtained a score that is indicative of a negative CA.

Regarding speech and language skills, all cautions were taken in order to exclude fluency difficulties from the calculation of the BVL scores. For instance, in the Naming and in the Sentence Completion tasks, children were given a wait time of up to 10 seconds before responding (unlike children with normal fluency who are given 10 seconds in total). The performance was considered valid if the subject, despite the disfluencies, completed the response within 10 seconds from starting. In the Narrative Fluency task and in the Multilevel Assessment of Narrative Speech task, we excluded Stuttering-like Disfluencies (SLD), such as syllable repetitions, in the count of units produced. In narrative fluency and in all other tests with time limits (Semantic Fluency and Phonological Fluency), we tried to be flexible with the imposed limit, granting the child additional time in case he lost time due to SLDs.

As to the results at the BVL battery, a detailed exposition of the individual performances at each sub-test is beyond the scope of the present paper. We summarise the results in table 1, where, for each sub-test, the number of children who obtained a performance to be classified in five performance ranges was calculated. These are: below -2 SD (deficit level), between -2 and -1 SD (lower than normal level), between -1 and +1 SD (level corresponding to the norm), between +1 and +2 SD (above the normal level) and above +2SD (far above the norm level).

Considering all the tests of the BVL battery, most of the participants (around 70 %) have obtained scores that fall within the average expected for their age (between -1 SD and +1 SD). The second following level is "lower than normal" (13.09 %), followed by "upper than normal" (8.20 %), then deficit level (7.34 %), and lastly, "far above the norm level" (1.27 %). Summing up the two levels under the norm level, we obtain around 20 % of all the children, while summing up the two levels above the norm, we obtain around 10 % of all the children.

The individual tests where most of the children scored less than -1 DS were (in percentages out of the total sample): Repetition of sentences(43.8 %); Repetition of words (44.4 %); Grammatical judgement (38.9 %); Articulation (naming) (33.3 %,

with 4 children scoring below -2 SD and 2 children scoring between -1 SD and -2 SD); Non-words repetition (33.3 %, with 3 children scoring below -2 SD and 3 children scoring between -1 SD and -2 SD).

Table 1 - *No. of children in each of the five performance levels considered for each test of BVL 4-12*

	< -2	*-2≤DS<-1*	*-1≤DS ≤1*	*1<DS≤2*	*> 2*
PRODUCTION					
Articulation	4	2	11	1	0
Naming	0	1	13	4	0
Semantic Fluency	0	1	12	3	2
Phonologic Fluency	2	3	9	2	0
Sentence Completion	1	1	14	1	0
No. of produced Words	0	2	14	1	1
Narrative Fluency	1	5	10	2	0
MLU	2	1	13	2	0
COMPREHENSION					
Phonologic Discrimination	0	1	17	0	0
Lexical Comprehension	0	2	13	1	0
Grammatical comprehension	0	1	15	2	0
Grammatical Judgement	2	5	10	1	0
Idiomatic Expressions	1	3	10	4	0
Linguistic Prosody	1	1	14	1	1
Affective Prosody	0	0	18	0	0
REPETITION					
Word Repetition	2	6	10	0	0
Non-Word Repetition	3	3	12	0	0
Sentence Repetition	4	3	8	1	0
ALL THE TASKS	23 (7.34 %)	41 (13.09 %)	223 (70.34 %)	26 (8.20 %)	4 (1.27 %)

Looking at the results of the three repetition tasks, the only one subject performing above mean level is outnumbered by the many subjects performing below the mean level, with substantial percentages scoring below two SDs, which is considered of clinical interest. Statistical investigations of correlation between the various performances at the individual test of the BVL 4-12 were conducted. The correlation index *Rho* for Spearman's ranks was used, a non-parametric statistical measure of correlation suitable for small samples, which do not follow a normal distribution, as in the case of the present study. The results of the analyses highlighted:

- A positive correlation between Semantic Fluency and Narrative Fluency (*Rho*=0.57, p<0.05);
- Positive correlations between Non-Word Repetition and: Grammatical Judgment (*Rho*=0.55, p<0.05), Articulation (*Rho*=0.60, p<0.01); Phonological discrimination (*Rho*=0.69, p<0.01), total mean of the production tests (*Rho*=0.63, p<0.01), total mean of all the other linguistic tests of the battery (*Rho*=0.78, p< 0.001);
- Positive correlations between the Phonological Discrimination and: Non-Word Repetition (*Rho*=0.69, p<0.01), Narrative Fluency (*Rho*=0.69, p<0.05), Articulation (*Rho*=0.54, p<0.05), total mean of the production tests (*Rho*=0.67, p<0.01), total mean of all the other linguistic tests of the battery (*Rho*=0.64, p<0.01);
- A positive correlation between the total mean of the production tests and the total mean of the comprehension tests (*Rho*=0.74, p<0.001).

No significant association was found between the individual scores, expressed in zeta-scores and averaged from all BVL 4-12 tests, and the individual raw stuttering severity scores from SSI-4 (*Rho*= -0.07, p=0.722), nor between the individual scores, expressed in zeta-scores and averaged from all BVL 4-12 tests, and the individual scores obtained from the KiddyCAT (*Rho*=-0.07, p=0.79). We also tested the degree of correlation between the individual raw stuttering severity scores from SSI-4 and the individual zeta-scores on every single test of the BVL 4-12. Although most of the correlations did not show statistical significance, there were two notable exceptions. For the BVL 4-12 Naming test, a remarkably significant and strong negative correlation emerged (*Rho*=-0.571, *p*=0.005). This indicates that as scores on the naming test increase, stuttering severity tend to decrease monotonically. Second, the Semantic Fluency test showed a negative correlation with SSI-4 scores (*Rho*=-0.481, p=0.023), suggesting that an increase in Semantic Fluency scores is associated with a decrease in stuttering severity.

Stuttering severity, as measured by both the individual raw scores of SSI-4 and by individual scores derived from the frequency subscale of dysfluencies, also did not correlate significantly with individual CA scores.

Finally, there were no statistically significant correlations between KiddyCAT scores and a) chronological age, b) age at stuttering onset, c) number of months elapsed since stuttering onset.

4. *Discussion and conclusions*

Our results turn out to be in line with the current literature, showing a substantial similarity between the speech and language skills of stuttering and nonstuttering peers, except for the very bad performances (<-2 SD) in the articulation task and in the word and non-word repetition tasks. The present study did not find any significant correlation between CA and stuttering severity (as in Winter, Byrd, 2021), CA and speech and language skills, CA and time of onset, or CA and time

elapsed from the stuttering onset to the experimental session. In our opinion, the absence of correlations between CA and stuttering severity could be explained by the unusual positive communicative attitude held by most of the subjects, with a mean score below the mean for CWS in Vanryckeghem *et al.* (2022), and by the absence of any certified comorbidity with other speech and language disorders. Here we provisionally advance an hypothesis inspired by the the results of Choi, Hurd & Kim (2022) on the level of awareness of disfluent speech in CWNS, that positive CA may be a consequence of not having a speech-language monitoring system sufficiently developed, causing most of CWS not to be aware of their problems and consequently not to be negatively affected in their CA.

The unexpected result that the most severe CWS obtained the lowest scores on the BVL 4-12 Naming subtest and the Semantic Fluency subtest seems to suggest for these subjects a possible problem in the lexical access process, but until it is possible to exclude a possible methodological bias related to the difficulty of the subjects more disfluent in respecting the time limits of these tasks (and to the related difficulty for the present authors to compensate for this), we prefer to suspend any further consideration.

Our final consideration involves the subjects' poor performance in repetition tasks, particularly NWR, regardless of severity scores. This result is in line with other studies (Pelczarski, Yaruss, 2016; Sasisekaran, Weathers, 2019) and it is consistent with the evidence of the high comorbidity of stuttering with Speech Sound Disorders (SSD). Interestingly, Gerwin *et al.* (2022) found that CWS with SSD perform significantly lower than CWS without SSD in a task of non-word repetition (NWR). NWR has also been claimed by Graf-Estes *et al.* (2007) and Coady, Evans (2008) as a powerful index for SLI. Regrettably, bad performance on NWR could be due to different causes (Krishnan *et al.*, 2017), because NWR reflects a variety of speech and language skills, including lexical access (e.g. Rubenstein *et al.* 1970), motor planning (e.g. Yoss, Darley 1974), phonological processing (e.g. Snowling 1981), phonological memory (e.g. Gathercole, Baddeley, 1989; Bowers, Bowers, Hudock & Ramsdell-Hudock, 2018), and the task also intercepts the processes of linguistic perception, phonological assembly and articulatory behaviour (see the review of Anderson *et al.*, 2019 pointing to CWS' weakness in executive functions; see also Sasisekaran, Basu, 2017). However, new light has been shed on NWR by Krishnan *et al.* (2017), claiming this task in children to be the reflection of the ability to compute and perform the sensorimotor conversions necessary to produce new words. This explanation matches the theories which interpret stuttering as a consequence of sensory-motor integration deficits (Speech Motor Control models based on efferent copy or hybrid feedback: Max *et al.*, 2004; Hickock *et al.*, 2011; Tian, Poeppel, 2012; Civier *et al.*, 2013; for a review, see Bradshaw e al., 2021). In fact, Bowers *et al.* (2018: 94) state that "cognitive-behavioral differences in NWR and dual-tasks are consistent with the involvement of neurocognitive networks related to executive function and sensorimotor integration in phonological working memory", and "mistimed or 'noisy' dorsal stream sensorimotor integration

previously proposed to cause stutter-typical disfluencies (i.e., part-word repetitions, prolongations, and blocks) may also account for load dependent differences on nonword repetition tasks in CWS and AWS". Last, but not least, these results could have a precious clinical value: Walsh *et al.* (2021) argued that prognostic index of chronic stuttering are: familiarity, low scores on standardized phonetic-phonological assessment tests, high frequency of SLD in spontaneous speech, and *low scores on NWR* (italics are of the present authors), and a prognostic value is explicitly assigned to NWR by Spencer, Weber-Fox (2014), Leech *et al.* (2017) and Mohan, Weber (2015).

We conclude by admitting that a limitation of the present research is the absence of a control group made up of CWNS matched by age and SES. This limit is to be understood as a suggestion for a better continuation of the project. Another proposal for the future is to increasingly expand the group of participants and diversify it, also recruiting stuttering subjects with certified comorbidity with language and/or speech disorders. A further extension of this study could concern ascertaining the status of persistent stuttering years after the performances of the subject of this contribution, and then using a statistical regression model to verify how the clinical outcome can predict the scores of linguistic proficiency and CA that we have presented here. This would help to clarify some results which are currently difficult to interpret, because they are potentially the result of two different populations: those of stutterers who will recover and those who will persist. That the two populations are distinguishable on the basis of behavioral data is a fact now shared by the majority of scholars (see Bloodstein, Bernstein-Ratner & Brundage, 2021; Yairi, Seery, 2023), and much of the current research on stuttering is aimed at identifying children predisposed to persistence early on the basis of some behavioral measures, which can then be used in a clinical setting. In fact, stuttering is characterized by having a very early average age of onset (33 months), but also a very high spontaneous recovery rate (around 90 %), which occurs, in 95 % of children who had begun to stutter, within the first 4 years of onset (Yairi, Ambrose, 2005). In light of these data, most of the CWS examined in this study could therefore currently be recovered, since when they were examined they had already been stuttering for many months (especially the older ones) and many more months have passed from the day of recording to date.

Bibliography

Adams, M.R. (1990). The demands and capacities model I: Theoretical elaborations. In *Journal of Fluency Disorders*, 15(3), 135-141.

Ambrose, N.G., Yairi, E. (1994). The development of awareness of stuttering in pre-school children. In *Journal of Fluency Disorders*, 19(4), 229-245.

Anderson, J.D., Conture, E.G. (2000). Language abilities of children who stutter: A preliminary study. In *Journal of Fluency Disorders*, 25(4), 283-304.

Anderson, J.D., Pellowski, M.W. & Conture, E.G. (2005). Childhood stuttering and dissociations across linguistic domains. In *Journal of Fluency Disorders*, 30(3), 219-253.

Anderson, J.D., Wagovich, S.A. & Hall, N.E. (2006). Nonword repetition skills in young children who do and do not stutter. In *Journal of Fluency Disorders*, 31(3), 177-199.

Anderson, J.D., Wagovich, S.A. (2010). Relationships among linguistic processing speed, phonological working memory, and attention in children who stutter. In *Journal of Fluency Disorders*, 35(3), 216-234.

Anderson, J.D., Wagovich, S.A. & Brown, B.T. (2019). Phonological and semantic contributions to verbal short-term memory in young children with developmental stuttering. In *Journal of Speech, Language, and Hearing Research*, 62(3), 644-667.

Arndt, J., Healey, E.C. (2001). Phonological neighborhood and word frequency effects in the stuttered disfluencies of children who stutter. In *Journal of Speech, Language, and Hearing Research*, 50, 229–247.

Bankson, N.W., Bernthal, J.E. (1990). *Bankson-Bernthal Test of Phonology*, USA: Applied Symbolix.

Bajaj, A. (2007). Working memory involvement in stuttering: Exploring the evidence and research implications. In *Journal of Fluency Disorders*, 32(3), 218-238.

Beilby, J.M., Byrnes, M.L. & Yaruss, J.S. (2012). The impact of a stuttering disorder on Western Australian children and adolescents. In *Perspectives on Fluency and Fluency Disorders*, 22(2), 51-62.

Bernardini, S., Vanryckeghem, M., Brutten, G.J., Cocco, L. & Zmarich, C. (2009). Communication attitude of Italian children who do and do not stutter. In *Journal of Communication Disorders*, 42(2), 155-161.

Bernardini, S., Cocco, L., Zmarich, C., Natarelli, G. & Vanryckeghem, M. (2019). The KiddyCAT: A normative and comparative preliminary study of italian CWS and CWNS. In Tomaiuoli D. (Ed.), *Proceeding of the 3rd International Conference on Stuttering, Rome, 14-16 June 2018*. Trento: Erickson, 241-248.

Boey, R.A., Van de Heyning, P.H., Wuyts, F.L., Heylen, L., Stoop, R. & De Bodt, M.S. (2009). Awareness and reactions of young stuttering children aged 2–7 years old towards their speech disfluency. In *Journal of Communication Disorders*, 42(5), 334-346.

Blood, G.W., Blood, I.M. (2004). Bullying in adolescents who stutter: Communicative competence and self-esteem. In *Contemporary Issues in Communication Science and Disorders*, 31(Spring), 69-79.

Blood, G.W., Blood, I.M., Tellis, G. & Gabel, R. (2001). Communication apprehension and self-perceived communication competence in adolescents who stutter. In *Journal of Fluency Disorders*, 26(3), 161-178.

Blood, G.W., Ridenour Jr, V.J., Qualls, C.D. & Hammer, C.S. (2003). Co-occurring Disorders in children who stutter. In *Journal of Communication Disorders*, 36(6), 427-448.

Bloodstein, O., Ratner, N.B. & Brundage, S.B. (2021). *A Handbook of Stuttering*. USA: Plural Publishing, Incorporated.

Bock, K., Levelt, W.J.M. (1994). Language Production: Grammatical Encoding. In Gernsbacher M.A. (Ed.), *Handbook of Psycholinguistics*. San Diego, CA: Academic Press, 945-984.

Bosshardt, H.G. (2006). Cognitive processing load as a determinant of stuttering: Summary of a research programme. In *Clinical Linguistics & Phonetics*, 20(5), 371-385.

Bowers, A., Bowers, L.M., Hudock, D. & Ramsdell-Hudock, H.L. (2018). Phonological working memory in developmental stuttering: potential insights from the neurobiology of language and cognition. In *Journal of Fluency Disorders*, 58, 94-117.

Bradshaw, A.R., Lametti, D.R. & McGettigan, C. (2021). The role of sensory feedback in developmental stuttering: a review. In *Neurobiology of Language*, 2(2), 308-334.

Brown, R. (1973). Development of the first language in the human species. In *American Psychologist*, 28(2), 97.

Vanryckeghem, M., Brutten, G.J. (2007). *KiddyCAT: Communication attitude test for preschool and kindergarten children who stutter*. San Diego, CA: Plural Publishing Incorporated.

Byrd, C.T., Conture, E.G. & Ohde, R.N. (2007). Phonological priming in young children who stutter: Holistic versus incremental processing. In *American Journal of Speech-Language Pathology*, 16, 43–53.

Chang, S.E., Angstadt, M., Chow, H.M., Etchell, A.C., Garnett, E.O., Choo, A.L., Kessler, D., Welsh, R.C. & Sripada, C. (2018). Anomalous network architecture of the resting brain in children who stutter. In *Journal of Fluency Disorders*, 55, 46-67.

Choi, D., Hurd, L.H. & Kim, H. (2022). Language abilities and awareness of stuttering in normally fluent children: preliminary findings. In *Speech, Language and Hearing*, 25(2), 225-234.

Chow, H.M., Garnett, E.O., Koenraads, S.P. & Chang, S.E. (2023). Brain developmental trajectories associated with childhood stuttering persistence and recovery. In *Developmental Cognitive Neuroscience*, 60, 103413.

Civier, O., Bullock, D., Max, L. & Guenther, F.H. (2013). Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. In *Brain and Language*, 126(3), 263-278.

Clark, C.E., Conture, E.G., Frankel, C.B. & Walden, T.A. (2012). Communicative and psychological dimensions of the KiddyCAT. In *Journal of Communication Disorders*, 45(3), 223-34.

Clark, C.E., Conture, E.G., Walden, T.A. & Lambert, W.E. (2015). Speech-language dissociations, distractibility, and childhood stuttering. In *American Journal of Speech-Language Pathology*, 24(3), 480-503.

Coady, J.A., Evans, J.L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). In *International journal of language & communication disorders*, 43(1), 1-40.

Conture, E.G., Zackheim, C.T., Anderson, J.D., Pellowski, M.W. & Ohde, R. (2004). Linguistic processes and childhood stuttering: Many'sa slip between intention and lip. In Maassen B., Kent R.D., Peters H.F.M., van Lieshout P.H.H.M. & Hulstijn W. (Eds.). *Speech motor control in normal and disordered speech*. Oxford, UK: Oxford University Press, 253–281.

Cook, S., Howell, P. (2014). Bullying in children and teenagers who stutter and the relation to self-esteem, social acceptance, and anxiety. In *Perspectives on Fluency and Fluency Disorders*, 24(2), 46-57.

Coulter, C.E., Anderson, J.D. & Conture, E.G. (2009). Childhood stuttering and dissociations across linguistic domains: A replication and extension. In *Journal of Fluency Disorders*, 34(4), 257-278.

Cream, A., Onslow, M., Packman, A. & Llewellyn, G. (2003). Protection from harm: the experience of adults after therapy with prolonged-speech. In *International Journal of Language & Communication Disorders*, 38(4), 379-395.

Daniels, D.E., Gabel, R.M. & Hughes, S. (2012). Recounting the K-12 school experiences of adults who stutter: A qualitative analysis. In *Journal of Fluency Disorders*, 37(2), 71-82.

De Nil, L.F., Brutten, G.J. (1990). Speech-associated attitudes. Stuttering, voice disordered, articulation Disorders, and normal speaking children. In *Journal of Fluency Disorders*, 15, 127-134.

Eriksson, M. (2017). The Swedish Communicative Development Inventory III: Parent reports on language in preschool children. In *International journal of behavioral development*, 41(5), 647-654.

Ezrati-Vinacour, R., Platzky, R. & Yairi, E. (2001). The young child's awareness of stuttering-like disfluency. In *Journal of Speech, Language and Hearing*, 44(2), 368-380.

Finneran, D.A., Francis, A.L. & Leonard, L.B. (2009). Sustained attention in children with specific language impairment (SLI). In *Journal of Speech, Language and Hearing Research*, 52(4), 915-29.

Gabel, R.M., Blood, G.W., Tellis, G.M. & Althouse, M.T. (2004). Measuring role entrapment of people who stutter. In *Journal of Fluency Disorders*, 29(1), 27-49.

Gathercole, S.E. (2006). Nonword repetition and word learning: The nature of the relationship. In *Applied Psycholinguistics*, 27(4), 513-543.

Gathercole, S.E., Frankish, C.R., Pickering, S.J. & Peaker, S. (1999). Phonotactic influences on short-term memory. In *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 84.

Gerwin, K.L., Walsh, B. & Tichenor, S.E. (2022). Nonword repetition performance differentiates children who stutter with and without concomitant speech sound and developmental language disorders. In *Journal of Speech, Language, and Hearing Research*, 65(1), 96-108.

Ginsberg, A.P., Wexler, K.B. (2000). Understanding stuttering and counseling clients who stutter. In *Journal of Mental Health Counseling*, 22(3), 228-239.

Estes, K.G., Evans, J.L., Alibali, M.W. & Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. In *Psychological Science*, 18(3), 254-260.

Gregg, B.A., Yairi, E. (2007). Phonological skills and disfluency levels in preschool children who stutter. In *Journal of communication Disorders*, 40(2), 97-115.

Groner, S., Walden, T., Jones, R. & Mary, J. (2016). Factors associated with negative attitudes toward speaking in preschool-age children who do and do not stutter. In *Contemporary Issues in Communication Science and Disorders*, 43, 255-267.

Guttormsen, L.S., Kefalianos, E. & Naess, K.A.B., (2015). Communication attitudes in children who stutter: A meta-analytic review. In *Journal of Fluency Disorders*, 46, 1-14.

Hakim, H.B. & Ratner, N.B. (2004). Nonword repetition abilities of children who stutter: An exploratory study. In *Journal of Fluency Disorders*, 29(3), 179-199.

Hall, N.E. (2004). Lexical development and retrieval in treating children who stutter. In *Language, Speech, And Hearing Services in Schools*, 35, 57–69.

Havstam, C., Sandberd, A.D. & Lohmander, A. (2011). Communication Attitude and Speech in 10-year-old children with cleft (and lip) palate: An ICF perspective. In *International Journal of Speech-Language Pathology*, 13, 156-164.

Hickok, G., Houde, J. & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. In *Neuron*, 69(3), 407-422.

Hoff, E., Tian, C. (2005). Socioeconomic status and cultural influences on language. In *Journal of Communication Disorders*, 38(4), 271-278.

Hollister, J., Van Horne, A.O. & Zebrowski, P. (2017). The relationship between grammatical development and disfluencies in preschool children who stutter and those who recover. In *American Journal of Speech-Language Pathology*, 26(1), 44-56.

Howell, P., Au-Yeung, J. (1995). The association between stuttering, Brown's factors, and phonological categories in child stutterers ranging in age between 2 and 12 years. In *Journal of Fluency Disorders*, 20(4), 331-344.

Howell, P., Au-Yeung, J. & Sackin, S. (2000). Internal structure of content words leading to lifespan differences in phonological difficulty in stuttering. In *Journal of Fluency Disorders*, 25(1), 1-20.

Johannisson, T.B., Wennerfeldt, S., Havstam, C., Naeslund, M., Jacobson, K. & Lohmander, A. (2009). The Communication Attitude Test (CAT-S): normative values for 220 Swedish children. In *International Journal of Language & Communication Disorders*, 44(6), 813-825.

Kawai, N., Healey, E.C., Nagasawa, T. & Vanryckeghem, M. (2012). Communication attitudes of Japanese school-age children who stutter. In *Journal of Communication Disorders*, 45(5), 348-54.

Kefalianos, E., Onslow, M., Packman, A., Vogel, A., Pezic, A., Mensah, F., Conway, L., Bavin, B., Block, S. & Reilly, S. (2017). The history of stuttering by 7 years of age: Follow-up of a prospective community cohort. In *Journal of Speech, Language, and Hearing Research*, 60(10), 2828-2839.

Krishnan, S., Alcock, K.J., Carey, D., Bergström, L., Karmiloff-Smith, A. & Dick, F. (2017). Fractionating nonword repetition: The contributions of short-term memory and oromotor praxis are different. In *PLoS One*, 12(7), e0178356.

Lescht, E., Venker, C., McHaney, J.R., Bohland, J.W. & Wray, A.H. (2022). Novel word recognition in childhood stuttering. In *Topics in Language Disorders*, 42(1), 41.

Louko, L.J., Edwards, M.L. & Conture, E.G. (1990). Phonological characteristics of young stutterers and their normally fluent peers: Preliminary observations. In *Journal of Fluency Disorders*, 15(4), 191-210.

Marini, A., Marotta, L., Bulgheroni, S. & Fabbro, F. (2015). *Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni*. Firenze, GIUNTI O.S.

McCormack, J., McLeod, S. & Crowe, K. (2019). What do children with speech sound disorders think about their talking?, *Seminars in Speech and Language*, 40, 2, 94-104.

Max, L., Guenther, F.H., Gracco, V.L., Ghosh, S.S. & Wallace, M.E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. In *Contemporary Issues in Communication Science and Disorders*, 31, 105-122.

Mohan, R., Weber, C. (2015). Neural systems mediating processing of sound units of language distinguish recovery versus persistence in stuttering. In *Journal of neurodevelopmental disorders*, 7, 1-13.

Montgomery, B.M., Fitch, J.L. (1988). The prevalence of stuttering in the hearing-impaired school age population. In *Journal of Speech and Hearing Disorders*, 53(2), 131-135.

Mulcahy, K., Hennessey, N., Beilby, J. & Byrnes, M. (2008). Social anxiety and the severity and typography of stuttering in adolescents. In *Journal of Fluency Disorders*, 33(4), 306-319.

Natke, U., Sandrieser, P., van Ark, M., Pietrowsky, R. & Kalveram, K.T. (2004). Linguistic stress, within-word position, and grammatical class in relation to early childhood stuttering. In *Journal of Fluency Disorders*, 29(2), 109-122.

Nippold, M.A. (2012). Stuttering and language ability in children: Questioning the connection, In *American Journal of Speech-Language Pathology*, 21, 183–196.

Nippold, M.A. (2019). Language development in children who stutter: A review of recent research. In *International Journal of Speech-Language Pathology*, 21(4), 368-376.

Nippold, M.A., Schwarz, I.E. & Jescheniak, J.D. (1991). Narrative ability in school-age stuttering boys: A preliminary investigation. In *Journal of Fluency Disorders*, 16(5-6), 289-308.

Ntourou K., Conture E.G. & Lipsey M.W. (2011). Language abilities of children who stutter: a Meta-Analytical Review. In *American Journal of Speech and Hearing Pathology*, 20, 3, 163-179.

Paden, E.P., Yairi, E. (1996). Phonological characteristics of children whose stuttering persisted or recovered. In *Journal of Speech, Language, and Hearing Research*, 39(5), 981-990.

Pelczarski, K.M., Yaruss, J.S. (2016). Phonological memory in young children who stutter. In *Journal of Communication Disorders*, 62, 54-66.

Perkins, W.H. (1990). What is stuttering?. In *Journal of Speech and Hearing Disorders*, 55(3), 370-382.

Pinton, A., Lena, L. (2015). *Fondamenti del trattamento logopedico in età evolutiva*. Italia: Carocci.

Quartarone, C., Bernardini, S., Bonichini, S., Natarelli, G., Pisciotta, C. & Zmarich, C. (2016). Verbal fluency, lexical production and socio-economic characteristics in a group of 24-months old children with a family history of stuttering, in relation to possible stuttering onset and final recovery or persistence. In D. Tomaiuoli (Ed.), *Proceeding of 2nd International Conference On Stuttering*, Rome, Italy – October 15-17, 2015, Trento: Erickson, 85-91 [ISBN: 978-88-590-1130-9].

Richels, C., Buhr, A., Conture, E. & Ntourou, K. (2010). Utterance complexity and stuttering on function words in preschool-age children who stutter. In *Journal of Fluency Disorders*, 35(3), 314-331.

Reilly, S., Onslow, M., Packman, A., Wake, M., Bavin, E.L., Prior, M., ... & Ukoumunne, O.C. (2009). Predicting stuttering onset by the age of 3 years: A prospective, community cohort study. In *Pediatrics*, 123(1), 270-277.

Riley, G. (2009). *The stuttering severity instrument for adults and children (SSI-4)*. Austin TX, Pro-Ed.

Roth, F.P., Spekman, N.J. (1986). Narrative discourse: Spontaneously generated stories of learning-disabled and normally achieving students. In *Journal of Speech and Hearing Disorders*, 51(1), 8-23.

Rubenstein, H., Garfield, L. & Millikan, J.A. (1970). Homographic entries in the internal lexicon. In *Journal of Verbal Learning and Verbal Behavior*, 9(5), 487-494.

Sasisekaran, J., Basu, S. (2017). The influence of executive functions on phonemic processing in children who do and do not stutter. In *Journal of Speech, Language, and Hearing Research*, 60(10), 2792-2807.

Sasisekaran, J., Weathers, E.J. (2019). Disfluencies and phonological revisions in a nonword repetition task in school-age children who stutter. In *Journal of Communication Disorders*, 81, 1-15.

Scott, L.A., Healey, E.C. & Norris, J.A. (1995). A comparison between children who stutter and their normally fluent peers on a story retelling task. In *Journal of Fluency Disorders*, 20(3), 279-292.

Seery, C.H., Watkins, R.V., Mangelsdorf, S.C. & Shigeto, A. (2007). Subtyping stuttering II: Contributions from language and temperament. In *Journal of Fluency Disorders*, 32(3), 197-217.

Shaw, D.M., Polikowsky, H.P., Pruett, D.G., Chen, H.H., Petty, L.E., Viljoen, K.Z., Beilby, J.M., Jones, R.M., Kraft, S.J. & Below, J.E. (2021). Phenome risk classification enables phenotypic imputation and gene discovery in developmental stuttering. *The American Journal of Human Genetics*, 108(12), 2271-2283.

Shimada, M., Toyomura, A., Fujii, T. & Minami, T. (2018). Children who stutter at 3 years of age: A community-based study. In *Journal of Fluency Disorders*, 56, 45-54.

Silverman, S., Ratner, N.B. (2002). Measuring lexical diversity in children who stutter: Application of vocd. In *Journal of Fluency Disorders*, 27(4), 289-304.

Smith, A., Weber, C. (2017). How stuttering develops: The multifactorial dynamic pathways theory. In *Journal of Speech, Language, and Hearing Research*, 60(9), 2483-2505.

Spencer, C., Weber-Fox, C. (2014). Preschool speech articulation and nonword repetition abilities may help predict eventual recovery or persistence of stuttering. In *Journal of Fluency Disorders*, 41, 32-46.

Snowling, M.J. (1981). Phonemic deficits in developmental dyslexia. In *Psychological Research*, 43(2), 219-234.

Starkweather, C.W., Gottwald, S.R. (1990). The demands and capacities model II: Clinical applications. *Journal of Fluency Disorders*, 15(3), 143-157.

Smith, A., Zelaznik, H. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. In *Developmental Psychobiology*, 45(1), 22-33.

Spaulding, T.J., Plante, E. & Vance, R. (2008). Sustained selective attention skills of preschool children with specific language impairment: Evidence for separate attentional capacities. In *Journal of Speech, Language, and Hearing Research*, 51(1), 16-34.

Subramanian, A., Yairi, E. & Amir, O. (2003). Second formant transitions in fluent speech of persistent and recovered preschool children who stutter. In *Journal of Communication Disorders*, 36(1), 59-75.

Tian, X., Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. In *Frontiers in Human Neuroscience*, 6, 314.

Unicomb, R., Kefalianos, E., Reilly, S., Cook, F. & Morgan, A. (2020). Prevalence and features of comorbid stuttering and speech sound disorder at age 4 years. In *Journal of Communication Disorders*, 84, 1-12.

Vanryckeghem, M., Hylebos, C., Brutten, G.J. & Peleman, M. (2001). The relationship between communication attitude and emotion of children who stutter. In *Journal of Fluency Disorders*, 26(1), 1-15.

Vanryckeghem, M., Brutten, G.J. (2007). *KiddyCat: Communication attitude test for preschool and kindergarten children who stutter*. San Diego, CA, Plural Publishing Incorporated.

Vanryckeghem, M., Brutten G.J. (2022), KiddyCAT – *Communication Attitude Test for Preschool and Kindergarten Children Who Stutter, Adattamento italiano di Simona Bernardini, Luisella Cocco e Claudio Zmarich*, Hogrefe.

Walden, T.A., Frankel, C.B., Buhr, A.P., Johnson, K.N., Conture, E.G. & Karrass, J.M. (2012). Dual diathesis-stressor model of emotional and linguistic contributions to developmental stuttering. In *Journal of Abnormal Child Psychology*, 40(4), 633-644.

Walsh, B., Mettel, K.M. & Smith, A. (2015). Speech motor planning and execution deficits in early childhood stuttering. In *Journal of Neurodevelopmental Disorders*, 7, 1-12.

Watkins, R.V., Yairi, E. & Ambrose, N.G. (1999). Early childhood stuttering III: Initial status of expressive language abilities. In *Journal of Speech, Language, and Hearing Research*, 42(5), 1125-1135.

Watts, A., Eadie, P., Block, S., Mensah, F. & Reilly, S. (2017). Language skills of children during the first 12 months after stuttering onset. In *Journal of Fluency Disorders*, n.

Weiss, A.L., Zebrowski, P.M. (1994). The narrative productions of children who stutter: A preliminary view. In *Journal of Fluency Disorders*, 19(1), 39-63.

Westby, C.E. (1979). Language performance of stuttering and nonstuttering children. In *Journal of Communication Disorders*, 12(2), 133-145.

Wijnen, F. (1990). The development of sentence planning. In *Journal of Child Language*, 17(3), 651-675.

Winters, K.L., Byrd, C.T. (2021). Predictors of communication attitude in preschool-age children who stutter. In *Journal of Communication Disorders*, 91, 1-10.

Yairi, E., Ambrose, N.G. (2005). *Early Childhood Stuttering for Clinicians by Clinicians*. Stati Uniti, PRO-ED.

Yairi, E., Seery C.H. (2023), *Stuttering: foundations and clinical applications*, San Diego, CA; Plural Publishing, Inc.

Yaruss, J.S., LaSalle, L.R. & Conture, E.G. (1998). Evaluating stuttering in young children: Diagnostic data. In *American Journal of Speech-Language Pathology*, 7(4), 62-76.

Yaruss, J.S., Quesal, R.W. (2004). Stuttering and the international classification of functioning, disability, and health (ICF): An update. In *Journal of Communication Disorders*, 37(1), 35-52.

Yoss, K.A., Darley, F.L. (1974). Developmental apraxia of speech in children with defective articulation. In *Journal of Speech and Hearing Research*, 17(3), 399-416.

Zackheim, C.T., Conture, E.G. (2003). Childhood stuttering and speech disfluencies in relation to children's mean length of utterance: A preliminary study. In *Journal of Fluency Disorders*, 28(2), 115-142.

Zaretsky, E., Lange, B.P., Euler, H.A., Robinson, F. & Neumann, K. (2017). Preschoolers who stutter score lower in verbal skills than their non-stuttering peers. In *The Buckingham Journal of Language and Linguistics*, 10, 96-115.

PARTE II

PARLATO NELLE DEMENZE E NELL'AUTISMO

SPEECH IN DEMENTIA AND AUTISM

DOI: 10.17469/O2111AISV000011

ELENA MARTINELLI, VITA GARRAMMONE, FRANCESCA MORI, INCORONATA NOLÈ, FRANCA CAMERIERO, MATILDE MARTINO, GAETANO DI BELLO, GLORIA GAGLIARDI

I correlati linguistici del deterioramento cognitivo: raccolta e analisi di un corpus di eloquio patologico di individui anziani lucani affetti da demenza

Linguistic correlates of cognitive decline: collection and analysis of a spoken corpus of patients with dementia living in Basilicata

The study aims to profile the linguistic-communicative disabilities related to dementia's onset in Italian. To this purpose, we recruited 40 subjects residing in Basilicata, an administrative region of Southern Italy not previously represented in this type of study. The enrolled cohort is balanced by sex and age: 20 healthy subjects and 20 subjects affected by different types of dementia (i.e., Alzheimer's disease, mixed dementia, frontotemporal dementia, and vascular dementia). After neuropsychological evaluation, we acquired approximately 9 hours of semi-spontaneous speech. Then, the corpus was transcribed and annotated to automatically extract a rich set of linguistic markers. Through statistical inference tests, we outline a degradation of the phonetic-acoustic, morphosyntactic/lexical, semantic, and syntax skills related to the disease. Finally, the qualitative analysis allows for better characterizing the linguistic profile of dementia, highlighting other traits such as local/global coherence and cohesion deficits, anomia's compensatory strategies, and repetitiveness/topic iteration.

Keywords: dementia, Italian, linguistic profiling, pathological speech, Clinical Linguistics.

1. *Caratteristiche linguistiche dell'eloquio nelle sindromi dementigene: lo stato dell'arte*

La demenza è una sindrome clinica caratterizzata dalla pervasiva compromissione delle funzioni cognitive e dalla conseguente perdita della personale autonomia funzionale. Si osserva in risposta all'insorgenza di un variegato spettro di patologie (es. malattie neurodegenerative, patologie cardiovascolari e/o metaboliche, ecc.) e costituisce la settima causa di morte a livello mondiale, la quarta nella popolazione ultrasettantacinquenne (Costa & Sinforiani, 2021).

Le stime epidemiologiche dell'incidenza e della mortalità riferite al disturbo nella popolazione anziana mostrano aumenti allarmanti: nel 2020 sono stati diagnosticati 50 milioni di casi, ne sono attesi 82 milioni nel 2030 e 152 milioni nel 2050 (ADI *et al*., 2021). Tali previsioni, in combinazione con il dispendio di risorse economiche, socio-assistenziali e sanitarie richieste, motivano la definizione della demenza come "priorità mondiale di salute pubblica" (OMS, 2012).

In virtù di quanto appena delineato, negli ultimi anni la comunità scientifica ha dedicato notevole attenzione all'identificazione di "marcatori" in grado di supportare la diagnosi, come *biomarker* liquorali, indici demografici e cardiovascolari. Più di recente, ha destato crescente interesse l'individuazione di *marker* linguistici della patologia: analizzare la produzione verbale di una coorte di pazienti e individuare le disabilità linguistico-comunicative associabili alla condizione morbosa sottesa rappresentano procedure estremamente vantaggiose (perché meno invasive e più economiche) per progettare strumenti che possano essere di ausilio alle pratiche diagnostiche, di prevenzione e riabilitazione clinica, nonché per migliorare le conoscenze relative al funzionamento della competenza linguistica normofasica (Gagliardi, 2019).

A partire dagli anni Settanta del Novecento, studi condotti a livello internazionale hanno permesso di evidenziare e descrivere il deterioramento delle competenze linguistiche causato dall'insorgenza di una patologia dementigena (cfr. Boschi *et al.*, 2017): alto numero di disfluenze, alterazione delle caratteristiche spettro-acustiche e ritmiche della voce, impoverimento morfo-sintattico e lessicale, difficoltà di *word-retrieval*, *deficit* di pianificazione ed ecolalie a livello pragmatico e testuale rappresentano gli aspetti linguistici deficitari tipicamente indotti dai processi neurodegenerativi (anche se i *pattern* sindromici esibiti e la gravità delle disfunzioni linguistiche variano a seconda dello specifico quadro dementigeno al quale il soggetto è ascrivibile (Papagno & Bolognini, 2020; Marino, 2021).

Costruire campioni rappresentativi della popolazione clinica oggetto di indagine costituisce un essenziale prerequisito per delineare il profilo linguistico-comunicativo di una malattia: se per alcune lingue si dispone di adeguati *corpora* di eloquio dementigeno (es. TalkBank – DementiaBank, per la lingua inglese; Becker *et al.*, 2004), le risorse per la lingua italiana sono estremamente limitate. Escludendo gli studi condotti su un numero piuttosto ristretto di pazienti, i *corpora* di parlato dementigeno raccolti per l'italiano sono pochi, a maggior ragione se si considera il fatto che non siano confluiti in raccolte pubblicate o accessibili. In aggiunta, la notevole variazione diatopica interna al sistema linguistico italiano (Berruto, 2021) impone che vengano acquisiti campioni di linguaggio reclutando soggetti appartenenti a quante più regioni della penisola italiana, affinché si possa fornire una solida base alle generalizzazioni che scaturiscono dalle analisi dei *corpora* raccolti.

I progetti più rilevanti in tale campo censiscono le produzioni linguistiche semi-spontanee di parlanti emiliano-romagnoli (progetto OPLON; Calzà *et al.*, 2021), campani (corpus CIPP-ma; Dovetto *et al.* 2022) e calabresi (*corpus* CIPP-mci; Dovetto *et al.* 2021). Al gruppo Anchise, invece, si deve la costruzione del primo *corpus* di parlato dementigeno spontaneo (Vigorelli, 2016; Benvenuti *et al.*, 2020): il *Corpus* Anchise 320, infatti, raccoglie le trascrizioni degli scambi conversazionali avvenuti in contesto ecologico tra operatori sanitari e soggetti affetti da demenza provenienti da molteplici regioni italiane (prevalentemente, però, dalla Lombardia).

2. *Razionale dello studio*

Lo studio si propone di confrontare a livello fonetico-acustico, morfosintattico, semantico, sintattico e pragmatico l'eloquio di soggetti italofoni affetti da demenza e quello di soggetti coetanei cognitivamente integri, al fine di evidenziare divergenze quantitative/qualitative funzionali alla profilazione delle caratteristiche linguistico-comunicative del deterioramento cognitivo per la lingua italiana.

Al fine di garantire una più affidabile generalizzabilità ai risultati ottenuti, nonché per disporre di materiale prezioso per futuri confronti in merito alla potenziale incidenza della variazione diatopica dell'italiano sui profili linguistici dei pazienti, sono stati reclutati soggetti nati e residenti in Basilicata, una regione non precedentemente rappresentata in studi di tale natura.

Il disegno sperimentale ha previsto l'acquisizione di campioni di linguaggio semi-spontaneo, appartenenti a due gruppi di soggetti adulti/anziani (> 60 anni), reclutati in seguito ad una valutazione neuropsicologica realizzata mediante la somministrazione di classici test carta e matita: *Mini-Mental State Examination* -MMSE (Magni *et al.*, 1996), *MOntreal Cognitive Assessment* – MoCA (Conti *et al.*, 2015) e due *test* di fluenza verbale, fonemica (Caltagirone *et al.*, 1995; Carlesimo *et al.*, 1995, 1996) e semantica (Spinnler & Tognoni, 1987).

Il campione risulta in definitiva costituito da 40 individui nati e residenti in Basilicata, di cui 24 di sesso femminile e 16 di sesso maschile, distinti in due gruppi (bilanciati per sesso ed età):

- Gruppo dei pazienti (GP): 20 soggetti (12 F, 8 M; età: 81 ± 6.9), con diagnosi conclamata di demenza e in cura presso la Residenza Sanitaria Assistenziale Opera Salute – Don Uva (sede di Potenza, PZ). Il gruppo è più finemente suddivisibile, sulla base della diagnosi ricevuta, in:
  - MA: 9 soggetti (8F, 1 M) con diagnosi di malattia di Alzheimer;
  - DM: 2 soggetti (1 F, 1M) con diagnosi di demenza mista;
  - DNS: 5 soggetti (2 F, 3 M) con diagnosi di demenza non ulteriormente specificata;
  - DV: 3 soggetti (M) con demenza vascolare;
  - DF-T: 1 soggetto (F) con demenza frontotemporale.
- Gruppo di controllo (GC): 20 soggetti (12 F e 8 M; età: 81 ± 6.3), con stato cognitivo nella norma.

Il gruppo dei pazienti è stato reclutato presso la Residenza Sanitaria Assistenziale "Universo Salute – Opera Don Uva" (PZ), durante un arco temporale di circa quattro mesi (gennaio 2022 – aprile 2022).

La successiva tabella riassume i criteri di reclutamento del campione.

Tabella 1 - *Criteri di reclutamento dei partecipanti allo studio*

CRITERI DI RECLUTAMENTO	
GRUPPO P	*GRUPPO C*
Età > 60 anni	Età > 60 anni
Esposizione linguistica monolingue	Esposizione linguistica monolingue
L1 italiano	L1 italiano
Diagnosi clinica di demenza	Assenza di deficit neurologici / sensoriali
Punteggi ai test carta-matita:	*Punteggi ai test carta-matita:*
– MMSE < 22	– MMSE ≥ 22
– MoCA ≤ 19,262	– MoCA > 19,262
– Fluenza fonemica < 17,35	– Fluenza fonemica ≥ 17,35
– Fluenza semantica < 7,25	– Fluenza semantica ≥ 7,25

Il progetto di ricerca è stato approvato dal Comitato di Bioetica dell'Alma Mater Studiorum, Università di Bologna (Prot. n. 72032 del 29/01/2022).

Per tutti i pazienti è stata ottenuta l'autorizzazione da parte dei relativi *caregiver*/tutori legali, in virtù del loro stato di deterioramento cognitivo; i soggetti di controllo, invece, hanno firmato personalmente i moduli di consenso informato per la partecipazione allo studio e di consenso per il trattamento dei dati personali.

I dati anagrafici, clinici e linguistici sono stati trattati nel rispetto della normativa vigente (Regolamento UE 2016/679 – Regolamento generale sulla protezione dei dati personali; D.Lgs. 196/2003 – Codice in materia di protezione dei dati personali): le informazioni richieste dal progetto e i risultati delle indagini sono stati annotati nella Scheda Raccolta Dati (CRF) e utilizzati in forma pseudo-anonimizzata attraverso l'attribuzione di un codice progressivo univoco ad ogni partecipante.

3. *Materiali e metodo*

3.1 Il corpus

L'eloquio dei partecipanti è stato acquisito mediante la somministrazione di tre *task* linguistici: due compiti narrativi, per i quali è stato richiesto – rispettivamente – il racconto di una gita o di viaggio compiuto nella propria vita (frase stimolo: "vuoi parlarmi di un viaggio o di una gita che hai fatto?") e quello delle tradizioni legate alla giornata di Natale (frase stimolo: "come passi di solito la giornata di Natale?"); un compito descrittivo, per il quale è stato impiegato come stimolo visivo una figura inclusa nel test neuropsicologico "Esame del Linguaggio II" (Ciurli *et al.*, 1996) (frase stimolo: "potresti descrivermi questo disegno?").

Per il gruppo di pazienti, le sessioni di registrazione sono state svolte nella RSA ospitante, quando possibile in stanze vuote/isolate al fine di limitare al minimo i rumori di sottofondo; le sessioni di registrazione dei soggetti del gruppo di controllo, invece, sono state svolte all'interno delle loro abitazioni. La somministrazione dei tre *task* ha richiesto in media circa 30 minuti.

Il *corpus* raccolto consta in totale di 8 h e 50 min di sonoro, di cui 7 h e 50 min acquisite per i gruppi reclutati e 1 h attribuibile alle produzioni dell'intervistatrice e di coloro che – in rari casi – hanno partecipato alle conversazioni anche se non direttamente intervistati (es. compagni di stanza dei pazienti sottoposti ad intervista nella RSA; familiari presenti durante le interviste dei soggetti di controllo).

3.2 Annotazione

Tutte le registrazioni sono state trascritte ortograficamente mediante il *software* ELAN (2021), secondo le convenzioni dell'Italiano Standard, impiegando il formato di trascrizione L-AcT (Cresti *et al.*, 2018).

Contestualmente alla trascrizione ortografica, è stata portata a termine la segmentazione dell'eloquio in enunciati, selezionati come unità di riferimento dell'analisi discorsiva in quanto concepibili – seguendo Austin (1962) – come minime entità linguistiche pragmaticamente interpretabili. I confini di enunciato sono stati annotati mediante l'identificazione dei *break* prosodici, indici percettivamente salienti poiché contraddistinti da una serie di caratteristiche acustiche, tra cui *reset* della F0, allungamenti finali e aumenti dell'intensità (cfr. Cresti & Moneglia, 2018).

La segmentazione prosodica del parlato acquisito e l'annotazione di alcuni fenomeni tipici della lingua parlata sono state dunque realizzate mediante una serie di diacritici proposti dal suddetto quadro teorico e brevemente specificati nella tabella 2.

Tabella 2 - *Diacritici – Framework teorico L-AcT*

Diacritico	*Spiegazione*
//	*break* prosodico terminale (conclusione di una sequenza discorsiva)
?	*break* prosodico terminale (conclusione di una sequenza discorsiva con profilo prosodico interrogativo)
/	*break* prosodico non terminale (conclusione di un'unità prosodica, ma non di un enunciato)
[/]	falsa partenza o un *retracting* con ripetizione
+	*break* prosodico coincidente con la terminazione non intenzionale dell'unità discorsiva progettata dal parlante, la cui comprensione risulta dunque potenzialmente compressa
#	esitazione temporanea silente o interruzione del flusso discorsivo (limite minimo: 250 ms)
&	parole incomplete e/o frammenti fonetici
< >	*overlapping* tra gli interattanti
xxx	parole incomprensibili
hhh	elementi paralinguistici / extralinguistici (es. risata, pianto)
yyy	parole non trascrivibili per *privacy*

La *pipeline* computazionale progettata da Gagliardi e Tamburini (2022) è stata successivamente impiegata al fine di estrarre i Biomarker Linguistici Digitali (*Digital Linguistic Biomarkers*, DLBs), *feature* linguistiche direttamente derivabili dalla produzione linguistica dei parlanti, la cui analisi consente di quantificare in maniera oggettiva la degradazione della facoltà di linguaggio in una data popolazione clinica (*infra*).

La procedura di estrazione di tali marcatori linguistici prevede alcune imprescindibili fasi di preprocessazione automatica dei dati, ovvero: segmentazione dell'eloquio acquisito in porzioni di silenzio e porzioni di parlato; tokenizzazione, lemmatizzazione, *PoS-tagging* e *parsing* sintattico a dipendenze, annotazioni realizzate sfruttando un *toolkit* di NLP *open source* (STANZA, v. 1.3.0; Qi *et al.*, 2020).

Prima di procedere con l'impiego della *pipeline*, gli *input* sono stati manualmente ripuliti: mediante il software *Audacity*, dai file audio sono state eliminate tutte le sezioni non riconducibili all'eloquio dei soggetti reclutati (interventi del ricercatore, rumori di sottofondo, sovrapposizioni tra interattanti) e la frequenza di campionamento di tutti i file è stata ridotta da 96 Khz / 24 bit a 44Khz / 16 bit (compatibile con la *pipeline*); dalle trascrizioni in formato .txt sono stati eliminati i diacritici precedentemente inseriti, ad eccezione dei *break* prosodici terminali ("//", sostituiti con il punto ".") e delle interruzioni improvvise degli enunciati ("+", sostituiti con i tre puntini sospensivi "...").

Inoltre, le espressioni inizialmente oscurate per *privacy* (= "yyy") sono state sostituite con un nome proprio (sempre lo stesso, scelto casualmente), al fine di continuare a tutelare la *privacy* dei soggetti senza intaccare la processazione delle produzioni linguistiche.

3.3 Analisi delle *feature* linguistiche

In prima istanza, la durata degli enunciati prodotti dai soggetti reclutati è stata valutata alla luce di una serie di variabili cliniche, ovvero sesso ed età dei partecipanti, presenza/assenza di declino cognitivo, tipologia e severità della demenza diagnosticata.

Successivamente, si è operato un confronto sulla ricca serie di DLBs estratti (tabella 3)[1], sia in relazione alla produzione linguistica complessiva, sia in relazione ai singoli *task*. Tale metodologia ha permesso dunque di individuare:

1. i correlati linguistici più "solidi" del deterioramento cognitivo, poiché significativamente differenti tra le due coorti nonostante l'eterogeneità dei *task* somministrati e la maggiore quantità di eloquio analizzato;
2. i "potenziali" correlati linguistici del deterioramento cognitivo, ovvero *feature* linguistiche meritevoli di futuri approfondimenti perché caratterizzanti l'eloquio dementigeno solamente in uno o due *task* ma non nella produzione linguistica complessiva.

[1] Per una definizione esaustiva dei molteplici DLBs estratti si rimanda a Calzà *et al.*, 2021.

La significatività delle divergenze riscontrate rispetto all'eloquio normofasico è stata analizzata mediante l'applicazione di *test* di inferenza statistica parametrici e non parametrici (es. T-test di Student, Test di Wilcoxon-Mann-Whitney) in R (R Core Team, 2014).

Tabella 3 - *Feature indagate*

DIGITAL LINGUISTIC BIOMARKERS (n°151)	
Livello fonetico-acustico (20 DLBs)	– Silence Segments Duration (Satt *et al.*, 2013) – Speech Segments Duration (Satt *et al.*, 2013) – Temporal Regularity of Voiced Segments (Satt *et al.*, 2013) – Verbal Rate (Singh *et al.*, 2001; Roark *et al.*, 2011) – Transformed Phonation Rate (Singh *et al.*, 2001; Roark *et al.*, 2011) – Standardized Phonation Time (Singh *et al.*, 2001; Roark *et al.*, 2011) – Standardized Pause Rate (Singh *et al.*, 2001; Roark *et al.*, 2011) – Root Mean Square Energy (López-de-Ipiña *et al.*, 2013) – Pitch (López-de-Ipiña *et al.*, 2013) – Spectral centroid (López-de-Ipiña *et al.*, 2013) – Higuchi Fractal Dimension (López-de-Ipiña *et al.*, 2013)
Livello lessicale (34 DLBs)	– Conten Density (Roark *et al.*, 2011) – Part-of-speech Rate (Bucks *et al.*, 2000) – Reference Rate to Reality (Vigorelli, 2004) – Personal, Spatial and Temporal Deixis Rate (March *et al.*, 2006; Cantos-Gòmez, 2009) – Relative pronouns and negative adverbs Rate – Lexical Richness: Type-Token Ratio (Holmes & Singh, 1996), Brunet's Index (Brunet, 1978) e R – Honor's Statistic (Honoré, 1979) – Action Verbs Rate (Gagliardi, 2014) – Frequency of use Tagging (De Mauro, 2000) – Propositional Idea Density (Snowdon *et al.*, 1996; Roark *et al.*, 2011) – Mean number of words for utterance
Livello semantico[2] *(84 DLBs)*	– Language – Emotiveness – Cognition – Sociality – Physicality – Lifestyle – Spatio-temporal references – Conversation style
Livello sintattico (9 DLBs)	– Number of dependent elements linked to the noun – Global dependency distance (Roark *et al.*, 2007; 2011) – Syntactic complexity (Szmrecsanyi, 2004) – Syntactic Embeddedness – Maximum Depth of The Structure – Utterance Length

[2] Tale gruppo di feature viene computato sulla base dell'analisi LIWC (*Linguistic Inquiry and Word Count*; cfr. Chung & Pennebaker, 2007). Nello specifico, le indagini sono basate sulla versione italiana del dizionario LIWC (Agosti & Rellini, 2007).

DIGITAL LINGUISTIC BIOMARKERS (n°151)	
Leggibilità[3] *(4 DLBs)*	– Basic readability – Lexical readability – Syntactic readability – Total readability

4. *Analisi quantitative: risultati*

4.1 Il profilo linguistico dei pazienti con deterioramento cognitivo

Le analisi condotte sui campioni di linguaggio raccolti hanno evidenziato un numero considerevole di divergenze: le successive tabelle espongono i dati statistici ottenuti analizzando la durata degli enunciati alla luce delle variabili cliniche oggetto di interesse (tabella 4), nonché indagando i valori dei molteplici tratti linguistici estratti tra i due gruppi reclutati (tabella 5).

Tabella 4 - *Durata degli enunciati: variabili cliniche*

Variabile clinica		*Test di inferenza statistica*		
		Mann-Whitney per campioni indipendenti	*Kruskall – Wallis*	*Correlazione di Kendall*
Diagnosi di demenza		< 0.001		
Tipo di demenza			< 0.001	
Severità della demenza	*MMSE*			< 0.001
	MOCA			< 0.001
	F_SE			< 0.001
	F_FO			< 0.001
Sesso	*GC*	0.08361		
	GP	< 0.001		
Età	*GC*			< 0.001
	GP			0.00333

[3] Tale *set* di *feature* viene direttamente computato grazie allo strumento di valutazione della leggibilità che, per la lingua italiana, attualmente è considerato il migliore sotto il profilo dell'affidabilità: READ-IT, sviluppato presso il CNR-ILC di Pisa (Dell'Orletta *et al.*, 2011).

Tabella 5 - *Digital Linguistic Biomarkers*

Livello	*DLB*	*T-test* / U-Test	
		p-value	*Significatività*[4]
Fonetico-acustico	SPE_SPEMEDIAN	0.001435	**
	SPE_TPR	0.001824	**
	SPE_SILMEDIAN	0.00304	**
	SPE_RMSEM	0.003717	**
	SPE_VR	0.004864	**
	SPE_SPEMEAN	0.006644	**
	SPE_SPCENTRM	0.009047	**
	SPE_SILSD	0.01121	*
	SPE_SILMEAN	0.02108	*
Lessicale	LEX_PoS_DET	0.0006802	***
	LEX_ContDens	0.0003114	***
	LEX_PoS_PUNCT	0.0002644	***
	LEX_PoS_ADJ	0.00006	***
	LEX_PoS_ADP	0.00006	***
	LEX_PoS__	1.1e-05	****
	LEX_RPRO	0.01136	***
	LEX_OCW	0.0002117	***
	LEX_CCW	0.0002117	***
	LEX_PoS_INTJ	0.00129	**
	LEX_TDEIXIS	0.001291	**
	LEX_PoS_AUX	0.002091	**
	LEX_PoS_X	0.002227	**
	LEX_ACTVRB	0.007714	**
	LEX_PoS_ADV	0.02447	*

[4] Livelli di significatività: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001, **** p-value < 0.0001.

Livello	*DLB*	*T-test* / U-Test	
		p-value	*Significatività*[4]
Semantico	LWC_Articol	0.03364	***
	LWC_Passato	0.0001192	***
	LWC_Prepos	0.0001554	***
	LWC_Tempo	0.0001854	***
	LWC_Ottimis	0.0004921	***
	LWC_Causa	0.0004983	***
	LWC_Sen_Pos	0.0007168	***
	LWC_Affett	0.001085	**
	LWC_Fisico	0.001609	**
	LWC_Noi_Verb	0.002149	**
	LWC_Sopra	0.004391	**
	LWC_Corpo	0.004105	**
	LWC_Scuola	0.004782	**
	LWC_Proc_Sen	0.007572	**
	LWC_P_pass	0.009773	**
	LWC_Consen	0.01018	*
	LWC_Tu_Verbo	0.01069	*
	LWC_Svago	0.01142	*
	LWC_Occupaz	0.02282	*
	LWC_Noi	0.02474	*
	LWC_Mangiare	0.02506	*
	LWC_Condizio	0.02594	*
	LWC_Movimen	0.02633	*
	LWC_riempiti	0.0263	*
	LWC_Sport	0.02854	*
	LWC_Raggiun	0.03484	*
	LWC_Rif_gen	0.03596	*
	LWC_Casa	0.03848	*
	LWC_Social	0.04381	*
Sintattico	SYN_MAXDEPTHM	0.0002984	***
	SYN_GRAPHDISTM	0.0004101	***
	SYN_MAXDEPTHSD	0.007158	**
	SYN_SLENM	0.01094	*
Leggibilità	REA_BASE	3,73E-06	****
	REA_LEXICAL	0.002643	**

Sintetizzando le evidenze riscontrate, si può affermare che l'eloquio dementigeno è caratterizzato da un numero maggiore di enunciati (tuttavia, significativamente più brevi) e da un numero maggiore di interruzioni improvvise della pianificazione discorsiva.

La severità della demenza, concepita in termini di punteggio ottenuto alle prove psicometriche somministrate, correla significativamente con la durata degli enunciati.

Per i pazienti, fattori anagrafici quali sesso ed età risultano significativamente influenti: nello specifico, la durata degli enunciati prodotti rivela una correlazione negativa con l'età e risulta divergere in base al sesso del soggetto. Poiché la condizione normotipica è rappresentata dal gruppo di controllo e, all'interno di tale popolazione, la focalizzazione sulle variabili anagrafiche non consente di evidenziare differenze statisticamente significative, si può ipotizzare che i risultati ottenuti per i pazienti siano in realtà specchio dell'azione di altri fattori: l'ineguale distribuzione dei tipi di demenza (e severità delle stesse) tra i pazienti dei due sessi, così come tra pazienti ascrivibili a una medesima fascia d'età, nonché possibili comorbidità di cui soffrono i soggetti anche quando ascrivibili a uno stesso quadro dementigeno.

Relativamente ai *Digital Linguistic Biomarker*, le indagini statistiche condotte sulla produzione linguistica complessiva dei soggetti permettono di affermare che l'eloquio patologico si differenzia da quello normotipico per 59/151 tratti, così suddivisibili: 9 DLBs acustici, 15 DLBs lessicali, 29 DLBs LWC, 4 DLBs sintattici, 2 DLBs di leggibilità.

Sintetizzando i risultati ottenuti, il parlato dementigeno risulta caratterizzato da una maggiore durata media e mediana dei segmenti di silenzio, nonché da una loro maggiore variabilità, e da una minore durata media e mediana dei segmenti di *speech*. La velocità di fonazione, riconducibile a un maggiore numero di pause, l'energia dei segmenti di *speech* e la *Verbal Rate* sono risultate significativamente minori.

Dal punto di vista lessicale, è osservabile – innanzitutto – una minore percentuale di parole di classe aperta: in particolare, emerge un ridotto impiego di aggettivi e avverbi, ma non di nomi e verbi, classi per le quali si osservano valori comparabili con quelli estratti per i soggetti controllo. La maggiore percentuale di parole di classe chiusa registrata è riconducibile prevalentemente a un maggiore ricorso alle interiezioni, poiché si osserva una minore occorrenza di preposizioni semplici e articolate, di determinanti, ma anche di avverbi interrogativi/congiunzioni subordinanti come "perché" (indice di un minore riferimento ai processi causativi). Senza dubbio, gli aspetti appena delineati concorrono a spiegare la minore densità di contenuto osservabile per l'eloquio dei pazienti.

Si evidenzia, ancora, una significativa riduzione delle espressioni deittiche di natura temporale, confermata da una generale riduzione delle espressioni temporali, soprattutto riferite/formulate al passato.

La maggiore percentuale di punteggiatura evidenziata dalle analisi, invece, è da interpretare come maggiore numero di enunciati e più frequenti interruzioni improvvise delle pianificazioni discorsive (*supra*). Questa *feature*, infatti, non va intesa come indicazione dell'uso dei segni di interpunzione (anche perché si tratta di testi orali), ma è da ricondursi alle modifiche testuali avvenute negli *step* di preprocessazione, durante i quali i *break* prosodici terminali ("//") e le interruzioni improvvise degli enunciati ("+") sono stati rispettivamente sostituiti dai segni di punteggiatura "." e "...".

Sebbene nel complesso non emergano differenze significative circa l'uso di verbi, l'eloquio dementigeno risulta essere comunque contraddistinto da una percentuale

minore di verbi d'azione e di verbi coniugati alla prima persona plurale, al netto di una maggiore percentuale di verbi coniugati alla seconda persona singolare. Minore è anche l'uso del modo condizionale, delle forme ausiliarie e del participio passato.

Lo stesso dicasi per l'uso dei pronomi: nel complesso le percentuali di occorrenza non differiscono, anche se si registra un minore impiego di forme pronominali relative e di forme pronominali di prima persona plurale.

L'analisi semantica degli enunciati prodotti dai pazienti ha evidenziato aspetti psicopatologici classicamente associati alle sindromi demenziali. Nello specifico, si osserva un generale decremento del riferimento alla sfera affettiva, dell'espressione di sensazioni positive e di atteggiamenti propositivi e consensuali. Gli atti referenziali risultano complessivamente ridotti, soprattutto se concernenti la sfera della socialità, i processi cognitivi di natura sensoriale (e soprattutto visiva), le attività di moto e lo spazio circostante.

Buona parte delle tematiche connesse alle attività quotidiane (occupazione, scuola, svago, ambiente domestico e sport) viene verbalizzata in minore misura; ridotta è anche la tendenza a discutere relativamente alla fisicità in generale (e, nello specifico, al corpo) nonché ad attività strettamente connesse con la cura di sé, come il nutrirsi.

Lo stile conversazionale, infine, si caratterizza per un minore numero di riempitivi.

Si conferma la complessiva preservazione delle competenze sintattiche, seppure sia evidenziabile una minore lunghezza media dell'enunciato e, in generale, una minore complessità delle strutture sintagmatiche, al netto di una maggiore variabilità circa la massima profondità dell'*embedding*.

Proprio relativamente alla lunghezza media dell'*embedding* si registra un dato interessante, poiché per i pazienti è risultato significativamente maggiore. Sebbene in letteratura sia ampiamente riconosciuta la generale preservazione delle competenze sintattiche nei soggetti affetti da demenza (ovvero la produzione di strutture sintattiche corrette ma semplificate), anche solo intuitivamente appare poco verosimile che i pazienti siano in grado di formulare costruzioni con un maggiore grado di *embedding* rispetto ai soggetti cognitivamente integri. A meno che non si tratti semplicemente di un errore a livello del *parsing* sintattico automatico, l'informazione ottenuta potrebbe trovare chiarimento mediante un'analisi qualitativa di livello macrolinguistico: il fatto che si registri un maggiore livello di incassatura sintattica non implica che, anche quando formalmente ben costruite, le costruzioni esibiscano lo stesso grado di correttezza a livello pragmatico-discorsivo. *Deficit* tutt'altro che rari, infatti, investono la capacità di produrre testi coesi e coerenti (a livello locale e/o globale), da cui consegue l'inefficacia comunicativa delle produzioni dei pazienti. Come si vedrà più avanti (§ 5), tali compromissioni si riflettono anche sulla struttura sintattica delle enunciazioni.

I dati relativi alla leggibilità confermano sostanzialmente quanto emerso analizzando i singoli livelli linguistici: l'eloquio dementigeno si caratterizza per una maggiore leggibilità di base, da ricondurre fondamentalmente alla minore lunghezza media degli enunciati e alla produzione di parole più brevi, così come

per una minore leggibilità lessicale, imputabile al fatto che i testi prodotti siano maggiormente imperniati su lemmi appartenenti al lessico fondamentale, ad alta frequenza e ad alta disponibilità (i.e., maggiore ambiguità lessicale dovuta alla bassa specificità del lessico impiegato), ma anche alla presenza di false partenze, frammenti fonetici e parafasie di diversa tipologia.

Il profilo linguistico-comunicativo finora tracciato può plausibilmente arricchirsi di ulteriori caratteristiche: *feature* linguistiche emerse come significativamente divergenti solo in uno dei *task* (eventualmente anche in due), ma non nella produzione linguistica complessiva. In questa sede vengono dunque presentati come "potenziali" correlati linguistici della demenza, poiché certamente meritevoli di futuri approfondimenti.

Da un punto di vista fonetico-acustico, si osserva una minore *Pause Rate* (indice di un maggiore numero di pause rispetto al numero di parole prodotte), una maggiore variabilità del Centroide Spettrale, così come un valore maggiore della *Higuchi Fractal Dimension* (indicativa del grado di somiglianza/identità tra strutture contigue nelle bande spettrali computate su sequenze temporali discrete) e una minore variabilità di tale indice.

Dal punto di vista lessicale, la lunghezza media dell'enunciato e le occorrenze di congiunzioni subordinanti, del verbo essere, dei nomi, dei verbi transitivi e del tempo presente, nonché il peso medio della frequenza d'uso delle parole sono risultati significativamente minori. L'uso dei nomi propri, invece, è certamente da ridefinirsi in futuro, poiché i dati emersi sono contrastanti.

Sul piano semantico, l'eloquio dementigeno talvolta pare caratterizzarsi per una minore espressione di processi cognitivi in generale, di riferimenti alla possibilità, di emozioni negative (specialmente della rabbia), di riferimenti agli affetti familiari e alle situazioni inclusive. Significativamente maggiori risultano, invece, sia il riferimento a processi introspettivi che l'espressione di inibizione.

Relativamente al piano sintattico, si dimostra significativamente minore per le produzioni linguistiche dei pazienti un'ulteriore misura della complessità sintattica, computata in termini di elementi linguistici suscettibili di incrementare la profondità delle strutture (es. congiunzioni subordinanti, pronomi *wh*).

Quest'ultimo dato viene confermato dalla maggiore leggibilità sintattica dei testi orali prodotti dai pazienti, emersa per il *task* 2 e riconducibile alla minore complessità delle strutture realizzate (in termini di articolazione interna dei periodi e delle proposizioni) e alla minore lunghezza delle relazioni di dipendenza (lunghezza media delle dipendenze e media delle dipendenze massime).

4.2 Quadri dementigeni e correlati linguistici

Nel complesso, i *Digital Linguistic Biomarker* risultati significativamente differenti tra i pazienti e i coetanei cognitivamente integri possono essere interpretati come correlati linguistici "più forti" della demenza: sono emersi nonostante l'eterogeneità del gruppo dei pazienti e dunque, presumibilmente, rappresentano aspetti linguistici più comunemente oggetto di degradazione, indipendentemente

dal processo neurodegenerativo sottostante. Allo stesso tempo, è bene sottolineare che proprio tale composizione mista può aver oscurato le peculiarità linguistiche di ogni quadro dementigeno: nonostante i sottogruppi di pazienti (ad eccezione dei soggetti con morbo di Alzheimer) non siano numerosi, la breve analisi in merito alla significatività delle *feature* estratte ha permesso di evidenziare per ognuno di essi un determinato corollario di correlati linguistici. Tuttavia i risultati devono essere interpretati con grande cautela: non sono infatti generalizzabili e certamente necessitano di future conferme.

Innanzitutto, la durata degli enunciati varia sensibilmente a seconda del quadro dementigeno al quale il paziente è ascrivibile. Da questo punto di vista, un risultato particolarmente rilevante concerne l'esito del confronto tra i dati estratti per i soggetti con demenza non ulteriormente specificata e quelli estratti per tutti gli altri pazienti: l'assenza di significatività registrata unicamente nel raffronto con i casi di demenza mista potrebbe suggerire l'identificazione dei casi di demenza non specificata con forme di demenza mista non ufficialmente diagnosticate. Se validata da indagini future, tale procedura potrebbe rappresentare un interessante supporto alla diagnosi differenziale in campo clinico.

In secondo luogo, la produzione linguistica dei soggetti affetti da Alzheimer risulta caratterizzata dalla maggiore percentuale di *feature* significativamente differenti tra eloquio dementigeno ed eloquio normotipico (72 %); il 33 % distingue il gruppo di soggetti con demenza vascolare, il 21 % il gruppo di soggetti con demenza non specificata, il 7 % il gruppo di soggetti con demenza mista e il 5 % il soggetto con demenza fronto-temporale.

Sebbene la maggior percentuale di tratti linguistici significativi registrata per i soggetti con Morbo di Alzheimer possa riflettere il fatto che numerose *feature* proposte in letteratura per l'identificazione delle patologie dementigene siano state individuate proprio mediante l'analisi del linguaggio alzheimeriano, i risultati ottenuti evidenziano piuttosto la necessità di acquisire nuovi dati: non è probabilmente un caso che la percentuale di tratti significativi (e la significatività stessa del *p-value*) decresca parallelamente al decrescere del numero di informanti per ogni sottogruppo.

Gli aspetti forse più interessanti nell'ottica di una discriminazione diagnostica *language-based* riguardano, da un lato, l'evidenziazione di caratteristiche significativamente differenti tra le singole coorti di pazienti (es. la percentuale maggiore di interiezioni dei soggetti con demenza vascolare rispetto ai soggetti con demenza non specificata, così come la minore complessità sintattica delle produzioni linguistiche di quest'ultima coorte rispetto a quanto si osserva per il soggetto con demenza fronto-temporale) e, dall'altro, l'identificazione di tratti significativamente divergenti solo tra alcuni sottogruppi di pazienti (es. la percentuale di disfluenze e di congiunzioni coordinanti, in entrambi i casi maggiori per i soggetti con demenza vascolare rispetto ai soggetti con malattia di Alzheimer) ma non tra il gruppo di pazienti e quello di controllo.

5. *Analisi qualitative: risultati*

L'analisi linguistica qualitativa ha consentito l'osservazione di ulteriori peculiarità dell'eloquio dementigeno.

Particolarmente evidente è la compromissione pragmatico-discorsiva: agli usi frequentemente agrammaticali di pronomi e congiunzioni, ad esempio, possono ricondursi la coesione referenziale e la coesione locale/globale deficitarie dei testi orali prodotti dai pazienti.

Per quanto riguarda i pronomi, è ben nota in letteratura la tendenza dei soggetti con demenza a impiegarne in alta percentuale: la ragione va sicuramente ricercata nella maggiore accessibilità e nel minore dispendio articolatorio che caratterizzano le forme pronominali, aspetti che permettono al paziente di portare a termine un atto referenziale cognitivamente più "economico" e altrimenti di difficile attuazione in virtù dei forti *deficit* di denominazione. La criticità più evidente, però, è relata al fatto che la pronominalizzazione nell'eloquio dementigeno spesso non riveste la sua funzione canonica: nonostante le occorrenze possano essere numericamente simili a quelle dei soggetti sani, i pazienti spesso ricorrono ai pronomi senza che sia effettivamente introdotto un referente nel cotesto linguistico antecedente/successivo e/o senza che ne venga indicato uno nel contesto extralinguistico. La comprensione dei riferimenti pronominali, anche di natura deittica, risulta perciò spesso impossibile, come esemplificato dal seguente estratto:

(1) SOGG_02 – Task 1[5]
8. si mettono sempre con # tutte quelle schifezze / # &po mi vengono a prendere / # e che se no # quello che &li devono fare //

Lo stesso dicasi per l'uso delle congiunzioni, sebbene sia necessaria una puntualizzazione: è più verosimile che i pazienti riescano a costruire costruzioni paratattiche grammaticalmente ben formate piuttosto che strutture ipotattiche ugualmente ben costruite, in virtù della maggiore dispendiosità cognitiva di queste ultime[6]. Detto ciò, non è scontato che la funzione delle congiunzioni subordinanti, anche se inserite nel discorso e computate come indici di complessità sintattica, sia poi adeguatamente concretizzata, similmente a quanto si è osservato per l'uso dei pronomi: anche se ne manifesta la progettazione, infatti, è possibile che il paziente non completi la realizzazione della subordinazione. Sono diversi gli esempi di questo tipo, ovvero enunciati nei quali il soggetto enuclea una costruzione ipotattica, salvo poi lasciarla in sospeso:

[5] Il SOGG_02, una paziente affetta da morbo di Alzheimer, impiega sì forme pronominali (sia esplicite che sottintese), ma per nessuna di esse è stato specificato/indicato un referente che possa rendere pienamente comprensibile il messaggio formulato.

[6] Le occorrenze di congiunzioni coordinanti risultano essere significativamente maggiori rispetto a quelle delle congiunzioni subordinanti, sia nei testi orali prodotti dai pazienti (Test di Kruskall-Wallis; *p-value* < 0.001) che in quelli prodotti dai soggetti di controllo (Test di Kruskal-Wallis; *p-value* < 0.001).

(2) SOGG_03 – Task 2a
1. sono loro che # stanno # &tr [/1] trovando delle [/1] delle avvolte / nelle cose così / perché +

(3) SOGG_07 – Task 2
3. se devo andare a # tanto per dire / # io vengo da te # prima / # dopo dieci giorni nemmeno +

Nei casi in cui, invece, le strutture sintattiche siano formalmente corrette, tangenzialità, prolissità e/o deragliamento ideativo plausibilmente spiegano perché le enunciazioni dei pazienti appaiono sia inefficaci dal punto di visto comunicativo, sia caratterizzate da un maggior grado di *embedding* da un punto vista quantitativo (cfr. § 4): come esemplificato dai seguenti passaggi, spesso i contenuti proposizionali veicolati risultano totalmente tangenziali rispetto al *topic* della frase principale (e, in generale, del *task*), e sono ricchi di informazioni irrilevanti, implausibili e/o ridondanti.

(4) SOGG_10 – Task 3
10. questo è un attrezzo # del medioevo / # che veniva usato per stirare [/1] # tirare la luna / # una volta ca [/1] fatto il cardinamento della lana / # a per si [/1] si fanno [/2] delle [/1] delle a forma di palla / # dei rotoli a forma di palla / # e allora xxx più tesa è la [/1] corda e più vengono bene le palle a forma di [/3] a forma di palla //

(5) SOGG_10 – Task 2
14. eh / una volta lì / non c'erano # medici # non c'erano i cosiddetti medici di famiglia / # allora la &buonanim di mio padre che / da questo io ho compreso che poi # era molto affezionato ai figli / # e difatti ne ha avuto nove / due sono morti piccolini / # e all'epoca non c'erano le cure che ci sono oggi / # e allora si moriva facilmente / quando si era &ba [/1] bimbi //

Il SOGG_10, durante la somministrazione del task 3, si pronuncia nel modo sopra riportato durante la descrizione dei ferri per il lavoro a maglia, rappresentati nella figura usata per l'elicitazione linguistica. Per quanto non si ravvisino agrammaticalità sintattiche, l'enunciazione non può dirsi corretta dal punto di vista semantico-lessicale e pragmatico-discorsivo. Sebbene sia stata espressa la funzione dell'oggetto da identificare (i ferri), nel discorso figura un'espressione certamente non pertinente (probabilmente una forma di parafasia verbale, ovvero "cardinamento"), così come appare non del tutto appropriata la definizione dell'oggetto come "attrezzo del medioevo". Gli enunciati prodotti durante il *task* 2, in aggiunta, ben esemplificano il deragliamento semantico e la verbalizzazione di informazioni fondamentalmente non pertinenti rispetto al *topic* del *task* (le tradizioni legate alla giornata del Natale), nonostante la correttezza grammaticale e la complessità delle costruzioni sintattiche realizzate.

Terminando questa sezione, si sottolinea che ulteriori caratteristiche ben evidenziabili dalla lettura del *corpus* riguardano:

1. l'uso della ripetizione come "riempitivo defaticante", al quale i soggetti ricorrono per sopperire alle difficoltà di prosecuzione dello scambio dialogico e che, a ben vedere, rappresenta la manifestazione "minima" di un più generale fenomeno di iterazione argomentativa, più evidente nei casi di minore severità della malattia

e testimonianza dell'esacerbazione dell'andamento epicicloidale che caratterizza i testi orali anche in condizioni normotipiche;
2. i riflessi linguistici dei severi *deficit* di denominazione da cui i pazienti sono affetti, a volte esplicitamente dichiarati, alle volte compensati da strategie come la circonlocuzione o l'impiego di espressioni deittiche, così come le frequenti parafasie fonemiche/fonetiche, semantiche e neologistiche;
3. la degradazione della memoria autobiografica, che si esplica nella trasposizione diacronica del passato nel presente e nella misidentificazione di affetti familiari in oggetti/persone estranee/non animate. Se l'uso di pronomi e aggettivi personali e possessivi sembra rispecchiare una percezione più o meno salda della propria identità, per i pazienti la collocazione spazio-temporale delle proprie esperienze di vita e dei propri affetti è invece evidentemente deficitaria.

6. *Conclusioni*

Complessivamente, i risultati sono stati coerenti con quanto riportato da studi condotti su altre lingue, mentre alcune delle *feature* proposte non sono risultate statisticamente significative (es. la maggiore percentuale di pronomi, di espressioni deittiche spaziali, la minore ricchezza lessicale, cfr. Boschi *et al.*, 2017; Millington & Luz, 2021): ciò a testimonianza di quanto ogni profilazione dei *deficit* linguistico-comunicativi debba essere tracciata alla luce della specifica lingua impiegata dai parlanti.

L'indagine qualitativa del *corpus* concorre a dettagliare il profilo delle disabilità linguistico-comunicative associabili alla demenza, permettendo di evidenziare *deficit* a livello pragmatico-discorsivo non desumibili esclusivamente dai dati quantitativi computati (es. errori di coesione referenziale e di coerenza locale/globale), nonché ulteriori peculiarità, tra cui notevole presenza di parafasie e false partenze, attuazione di strategie in compensazione della severa anomia e contestualizzazione spazio-temporale degli eventi deficitaria.

Tutte le *feature* risultate già significative nel confronto tra le condizioni di *Mild Cognitive Impairment* (MCI; cfr. Petersen *et al.*, 1999) e di integrità cognitiva per soggetti italofoni (cfr. Calzà *et al.*, 2021) sono risultate altrettanto discriminanti in questo studio: fanno eccezione solo due indici di ricchezza lessicale (TTR e BrunetW). La motivazione potrebbe risiedere in errori di annotazione (e successiva computazione) automatica, o più semplicemente nelle differenti fasce d'età coinvolte nei due studi: con l'avanzare dell'età, la ricchezza lessicale decresce anche in condizioni di normalità cognitiva e, dunque, per i soggetti più anziani tali aspetti semantico-lessicali potrebbero non essere più significativamente differenzianti.

Un aspetto cruciale del disegno sperimentale presentato in questa sede concerne indubbiamente le modalità di annotazione del *corpus* acquisito e, conseguentemente, di estrazione dei DLBs: l'annotazione a tutti i livelli linguistici è avvenuta in maniera automatica, senza che – per questioni di tempo – si procedesse con una revisione manuale degli *output* prodotti dalla *pipeline*. Trattandosi di procedure automatiche, la percentuale di errore è certamente più alta rispetto a un'annotazione manuale,

ma la questione più spinosa è di natura sociolinguistica: sebbene i soggetti reclutati siano italofoni, sono tutt'altro che rare le inflessioni dialettali (a livello soprattutto lessicale); non è da escludere, dunque, che vi possano essere stati errori di *PoS-Tagging*, lemmatizzazione e, conseguentemente, di *parsing* sintattico, dovuti al mancato riconoscimento delle forme dialettali usate dai soggetti.

Nonostante queste problematiche, i risultati ottenuti sono molto promettenti, poiché la semplice annotazione automatica ha permesso di evidenziare un considerevole numero di tratti significativamente correlati alla demenza. Sarebbe certamente interessante ripetere le analisi condotte dopo aver revisionato manualmente l'annotazione automatica del *corpus*: da un lato, per individuare eventuali divergenze rispetto ai DLBs risultati significativi in questo studio e, dunque, per incrementare/diminuire i correlati linguistici della demenza per la lingua italiana; dall'altro, per avere una stima della bontà dei risultati ottenibili esclusivamente con una totale automatizzazione delle procedure di annotazione ed estrazione delle *feature* linguistiche.

In futuro si dimostrerà sicuramente fruttuoso ampliare il *corpus* già raccolto, reclutando un maggiore numero di pazienti: 1) ascrivibili a quadri dementigeni differenti, in modo da approfondirne le singole profilazioni linguistico-comunicative; 2) affetti da demenza di severità differente, così da osservare al meglio variazioni delle competenze linguistiche legate alla gravità della compromissione cognitiva; 3) con provenienza regionale differente, affinché si possano individuare anche eventuali divergenze relate alla ricca variazione diatopica della lingua italiana; 4) con vari livelli di istruzione, al fine di valutare la possibile persistenza di tratti linguistici primariamente imputabili all'utilizzo di varietà *substandard* dell'italiano e ad una considerevole esposizione a sistemi linguistici dialettali.

L'acquisizione di nuovi dati sarà naturalmente preziosa per la generalizzabilità dei risultati e per il miglioramento dei processi di annotazione linguistica automatica, poiché i sistemi di NLP implicati (es. ASR, *Pos-Tagging*, ecc.) esigono grandi quantità di dati relativi alle produzioni linguistiche non normotipo per essere addestrati e ottenere migliori prestazioni.

Infine, incrementare i *Digital Linguistic Biomarker* da indagare (per le più recenti rassegne al riguardo: de la Fuente Garcia *et al.*, 2020; Petti *et al.*, 2020; Vigo *et al.*, 2022) e integrare i risultati quantitativi con riflessioni qualitative consentirà indubbiamente di delineare una caratterizzazione linguistico-comunicativa del deterioramento cognitivo più precisa e completa.

CRediT Author Statement

EM: Investigation, Data Curation (i.e., linguistic samples collection), Writing – Original Draft; VG, FM, IN, FC, MM, GDB: Resources (i.e., patient enrollment); GG: Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Riferimenti bibliografici

Alzheimer's Disease International (2021). *World Alzheimer Report 2021: Journey through the diagnosis of dementia*. London: Alzheimer's Disease International.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington (DC): American Psychiatric Publishing. DOI: 10.1176/appi.books.9780890425596

Austin, J.L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J. & McGonigle, K.L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. In *Archives of neurology*, 51(6), 585-594. DOI: 10.1001/archneur.1994.00540180063015

Benvenuti, N., Bolioli, A., Mazzei, A., Vigorelli, P. & Bosca, A. (2020). The "Corpus Anchise 320" and the analysis of conversations between healthcare workers and people with dementia. In Dell'Orletta, F., Monti, J. & Tamburini, F. (a cura di.), *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*. Torino: Accademia University Press.

Berruto, G. (2021). *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci Editore.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A. & Cappa, S.F. (2017). Connected speech in neurodegenerative language disorders: a review. In *Frontiers in Psychology*, 8, 269. DOI: 10.3389/fpsyg.2017.00269

Caltagirone, C., Gainotti, G., Carlesimo, G.A. & Parnetti, L. (1995). Batteria per la valutazione del deterioramento mentale: I. Descrizione di uno strumento di diagnosi neuropsicologica. In *Archivio di Psicologia, Neurologia e Psichiatria*, 56(4), 461-470.

Calzà, L., Gagliardi, G., Rossini Favretti, R. & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying Mild Cognitive Impairment and dementia. In *Computer Speech & Language*, 65, 101-113. DOI: 10.1016/j.csl.2020.101113

Carlesimo, G., Caltagirone, C., Gainotti, G. & Nocentini, U. (1995). Batteria per la valutazione del deterioramento mentale: standardizzazione ed affidabilità diagnostica nell'identificazione di pazienti affetti da sindrome demenziale. In *Archivio di Psicologia Neurologia e Psichiatria*. 56, 471-488.

Ciurli, P., Marangolo, P. & Basso, A. (1996). *Esame del Linguaggio II. Manuale e materiale d'esame*. Firenze: Giunti.

Conti, S., Bonazzi, S., Laiacona, M., Masina, M. & Vanelli Coralli, M. (2015). Montreal Cognitive Assessment (MoCA) – Italian version: regression-based norms and and equivalent scores. In *Neurological Sciences*, 26, 209-214. DOI: 10.1007/s10072-014-1921-3

Costa, A., Sinforiani, E. (2020). *Le demenze. Manuale di diagnosi e trattamento*. Milano: NEMS.

Cresti, E., Moneglia, M. (2018). Chapter 13. The illocutionary basis of Information Structure: The Language into Act Theory (L-AcT). In Adamou, E. et al. (a cura di), *Information Structure in Lesser-described Languages: Studies in prosody and syntax*. Amstredam: John Benjamins Publishing Company, 360-402.

Crystal, D. (1981). *Clinical Linguistics*. Wien: Springer Verlag.

de la Fuente Garcia, S., Ritchie, C.W. & Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. In *Journal of Alzheimer's disease*, 78(4), 1547-1574. DOI: 10.3233/JAD-200888

De Mauro, T. (2000). *Il dizionario della lingua italiana*. Torino: Paravia.

Dell'Orletta, F., Montemagni, S. & Venturi, G. (2011). READ–IT: Assessing readability of Italian texts with a view to text simplification. In Norman, A. (a cura di), *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Stroudsburg (PA): ACL, 73-83.

Dovetto, F.M., Guida, A., Pagliaro, A.C., Guarasci, R., Raggio, L., Sorrentino, A. & Trillocco, S. (2022). Corpora di Italiano Parlato Patologico dell'età adulta e senile. In Cresti, E., Moneglia, M. (a cura di), *Corpora e Studi Linguistici. Atti del LIV Congresso della Società di Linguistica Italiana*. Milano: Officinaventuno, 165-177.

Max Planck Institute for Psycholinguistics (2021). ELAN. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [Computer Software] Version 6.2. https://archive.mpi.nl/tla/elan

Gagliardi, G., Tamburini, F. (2021). Linguistic biomarkers for the detection of Mild Cognitive Impairment. In *Lingue e linguaggio*, 1/2021, 3-31.

Gagliardi, G. (2019). *Linguistica per le professioni sanitarie*. Bologna: Pàtron.

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G. & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. In *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10(1), 260–268. DOI: 10.1016/j.dadm.2018.02.004

Honore, A. (1979). Some simple measures of richness of vocabulary. In *Association of Literary and Linguistic Computing Bulletin*, 7, 172-177.

Magni, E., Binetti, G., Bianchetti, A., Rozzini, R. & Trabucchi, M. (1996). Mini-Mental State Examination: a normative study in Italian elderly population. In *European Journal of Neurology*, 3(3), 198–202. DOI: 10.1111/j.1468-1331.1996.tb00423

Pagano, C., Bolognini, N. (a cura di) (2020). *Neuropsicologia delle demenze*. Bologna: Il Mulino.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G. & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. In *Archives of neurology*, 56(3), 303-308. DOI: 10.1001/archneur.56.3.303

Petti, U., Baker, S. & Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. In *Journal of the American Medical Informatics Association – JAMIA*, 27(11), 1784-1797. DOI: 10.1093/jamia/ocaa174

Qi, P., Dozat, T., Zhang, Y. & Manning, C.D. (2018). Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Stroudsburg (PA): ACL, 160-170. DOI: 10.18653/v1/K18-2016

R Core Team (2014). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical.

Roark, B., Mitchell, M. & Hollingshead, K. (2007). Syntactic complexity measures for detecting Mild Cognitive Impairment. In Bretonnel Cohen, K., Demner-Fushman, D., Friedman, C., Hirschman, L. & Pestian, J. (a cura di), *Proceedings of the Workshop BioNLP 2007: Biological, translational, and clinical language processing*. Stroudsburg (PA): ACL, 1-8.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. & Kaye, J.A. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. In *IEEE Transactions on Audio Speech, and Language Processing*, 19(7), 2081-2090. DOI: 10.1109/TASL.2011.2112351

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A. & Tsolaki, M. (2013). Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia. In *Proceedings of Interspeech 2013*. Grenoble: ISCA, 1692–1696.

Spinnler, H., Tognoni, G. (1987). Standardizzazione e taratura italiana di test neuropsicologici. Milano: Masson Italia periodici *(Supplementum 8 – Italian journal of neurological sciences)*.

Vigo, I., Coelho, L. & Reis, S. (2022). Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review. In *Bioengineering*, 9(1):27. DOI: 10.3390/bioengineering9010027

World Health Organization (2012). *Dementia: a public health priority*. Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/75263

DOI: 10.17469/O2111AISV000012

FRANCESCO SIGONA, BARBARA GILI FIVELA, PIETRO VIGORELLI, ANDREA BOLIOLI

A computational linguistic analysis of speech in elderly individuals with cognitive decline: the "Anchise-2022"

In this article, we introduce the "Anchise-2022" corpus, the latest update of the "Anchise" corpus, with approximately 600 manually transcribed conversations, mainly between healthcare providers and elderly individuals with cognitive decline (dementia, Alzheimer's disease). We also analysed over 200 of these transcripts, to check if and to what extent we can correlate characteristics of such kind of speech production with the progression of dementia as measured by Mini-Mental State Examination (MMSE). To this purpose, we adopted computational linguistic analysis using Natural Language Processing (NLP) techniques, statistical-descriptive and inferential analysis. The results obtained allow us to assert, in line with existing literature, that certain parameters exhibit a degree of correlation with MMSE values. Additionally, it is feasible to utilize automatic classification techniques to infer the degree of linguistic/cognitive impairment.

Keywords: MMSE, NLP, Linguistic features, Ecological conversations, Enabling Approach.

1. *Introduction*

The importance of detecting language impairments has grown significantly in the diagnosis of various neurodegenerative diseases. These language deficits can manifest in different ways within neurodegenerative conditions. They may appear as a prominent symptom in the early stages, as seen in Primary Progressive Aphasia, or occur alongside other cognitive issues, as seen in Alzheimer's Disease (AD) (Boschi, Catricalà, Consonni, Chesi, Moro & Cappa, 2017). Additionally, individuals with Mild Cognitive Impairment (MCI) who later develop Dementia also exhibit language deficits (Karr, Graham, Hofer & Muniz-Terrera, 2018). Studies analysing connected speech in individuals with MCI related to AD and cognitively healthy individuals have highlighted significant differences, primarily related to reduced fluency and semantic impoverishment (Filiou, Bier, Slegers, Houzé, Belchior & Brambati, 2020). Furthermore, research comparing individuals with progressive AD and those who do not progress has found significant disparities in naming ability and semantic fluency (Kim, B.S., Kim, Y.B. & Kim, H., 2019). Language difficulties pose a substantial hurdle for dementia patients, particularly as the disease advances. In the early stages, language impairments typically involve difficulties in word retrieval, reduced verbal fluency, and compromised comprehension of written and spoken language (Banovic, Zunic & Sinanovic, 2018). As AD progresses, verbal fluency further declines, comprehension diminishes, and literal

and semantic paraphrases become more prevalent. In advanced stages, speech often becomes limited to repetitive echoing (echolalia) and verbal stereotypes (Soria Lopez, González & Léger, 2019). Given that analysing speech production closely mirrors real-life situations and daily functioning, it holds promise for assessing cognitive abilities and disease progression. However, traditional standardized psychometric instruments for evaluating linguistic abilities are time-consuming and prone to human bias. Recent studies have pointed out issues with standardization, normative data, criterion validity, and reliability of tools developed for non-neurodegenerative language and communication impairments (Krein, Jeon, Miller Amberber & Fethney, 2019). Additionally, these tools often overlook the perspective of individuals living with dementia (ILWD) regarding the impact of language impairments on their daily lives. Automatic analysis of speech and language (AASL) using Natural Language Processing (NLP) techniques has gained traction for assessing language proficiency, cognitive abilities, and even motor skills when audio signals are available. Numerous studies in the literature have focused on text-based AASL, correlating its features with psychometric scores and distinguishing between healthy and unhealthy subjects (Vigo, Coelho & Reis, 2022). However, there's a lack of consensus regarding the choice of tasks, features, and algorithms due to insufficient information (Gagliardi, 2023). Furthermore, most existing studies involve controlled experiments with pre-established tasks, which introduce some level of structure into speech output and result in semi-spontaneous speech (Prins, Bastiaanse, 2004). Research on spontaneous speech of ILWD in everyday interactive communication settings with high ecological validity has been limited. Conversations in ecological settings offer valuable insights, as they reflect a more genuine expression compared to controlled experiments. The Anchise Corpus, consisting of face-to-face conversations between ILWD in nursing homes and caregivers using the Enabling Approach, provides a unique dataset characterized by high ecological validity. It offers a substantial number of samples across various stages of cognitive decline, unlike smaller samples in controlled trials, resulting an essential resource for our investigation. This paper aims to evaluate the effectiveness of statistical analyses and classification approaches of NLP feature extracted by NLP techniques from transcribed dialogues in highly ecological contexts, in characterizing the oral language of speakers with dementia at various stages and identifying correlations between speech transcripts and patients' MMSE scores.

2. *Related work*

2.1 Elicitation tasks and speech resources

Adequate speech corpora for individuals with cognitive impairments are available for certain languages, such as English (e.g., TalkBank – DementiaBank, Becker, Boiler, Lopez, Saxton & Mcgonigle, 2004; for a comprehensive list of speech databases, see Vigo *et al.*, 2022). However, resources for the Italian language are notably scarce in comparison. Often, these resources involve a limited number

of participants and have not been compiled into published or readily accessible collections. The primary initiatives in this regard, which focus on analysing semi-spontaneous speech in specific Italian dialects, include the OPLON project (Beltrami, Gagliardi, Rossini Favretti, Ghidoni, Tamburini & Calzà, 2018; Calzà, Gagliardi, Rossini Favretti & Tamburini, 2021), the CIPP-ma corpus, and the CIPP-mci corpus (Dovetto, Guida, Pagliaro, Guardasci, Raggio, Sorrentino & Trillocco, 2022). These corpora have been assembled using well-established tasks. A frequently employed approach to collect speech samples involves having patients describe a scene, as demonstrated in the "Cookie Theft" picture (Goodglass, Kaplan & Weintraub, 1983). Traditionally, picture description has been considered more advantageous when compared to other discourse production tasks that have less structure, such as informal conversation or procedural tasks (Ulatowska, Allard, Donnell, Bristow, Haynes & Flower, 1988). Additionally, among various cognitively demanding tasks, recounting a story, whether it's a well-known tale like Pinocchio or Cinderella or a novel narrative presented through pictures (Toledo, Aluísio, Santos, Brucki, Trés, Oliveira *et al.*, 2017), requires the integration of characters and events within a temporal framework. Reading (with or without recall), narrating personal life experiences, dreams, and similar tasks are also commonly used. Petti, Baker & Korhonen (2020) and de la Fuente Garcia, Ritchie & Luz (2020) have provided comprehensive reviews of relevant studies in this area. Finally, in numerous instances, conversations between patients and doctors are documented during clinical assessments or cognitive evaluations. For example, in the study by Mirheidari, Blackburn, Walker, Reuber & Christensen (2019), neurologists were instructed to adhere to a predetermined set of questions explicitly designed to uncover common signs of impairments throughout the conversation. These questions included closed queries requiring long-term memory recall, compound questions, open-ended inquiries, and those related to memory concerns. Weiner and Schultz (2016) conducted an analysis of semi-standardized biographical interviews within the ILSE corpus (Weiner, Frankenberg, Telaar, Wendelstein, Schröder & Schultz, 2016). Espinoza-Cuadros, Garcia-Zamora, Torres-Boza, Ferrer-Riesgo, Montero-Benavides, Gonzalez-Moreira & Hernandez-Gomez (2014) recorded speech interactions from both clinicians and patients during structured interviews used in the administration of the Spanish version of the MMSE. Additionally, Luz, de la Fuente & Albert (2018) examined natural conversations taken from the Carolina Conversations Collection (Pope, Davis, 2011), which involved 21 ILWD and 17 individuals with no neuropsychological diseases (interviewers included gerontologists, linguistic students, or researchers).

2.2 Psychometric-clinical indices

There are various methods and scales available to assess the severity of cognitive impairments, dementia, and Alzheimer's disease (AD). The choice of method or scale depends on the specific context, whether it's research or clinical practice,

and the aspect being evaluated. For measuring overall outcomes, some commonly used scales include the Global Deterioration Scale (GDS) and the Clinical Global Impression of Change (CGI-C). To assess functional ability and quality of life, the Progressive Deterioration Scale (PDS) and the Disability Assessment for Dementia (DAD) are often employed. Cognitive ability is typically evaluated using various tools such as the Alzheimer's Disease Assessment Scale – cognitive subscale (ADAS-cog), the Severe Impairment Battery (SIB), the Montreal Cognitive Assessment (MoCA), and the Mini-Mental State Examination (MMSE, Folstein, M., Folstein, S & Mchugh, 1975). The Mini-Mental State Examination (MMSE) remains one of the most widely employed assessment tools today. It is a brief test, typically taking 10-15 minutes to complete, consisting of 30 items that assess various cognitive domains. These domains encompass temporal-spatial orientation, short-term memory, attention and calculation, the ability to recall memories, comprehension and language, nominal ability, as well as the ability to write sensibly and to replicate minimally complex designs. Studies have reported that the MMSE demonstrates internal consistency and exhibits short-term test-retest reliability in ILWD and long-term reliability in those with intact cognitive functioning. Additionally, it has proven to be sensitive to the severity of dementia in Alzheimer's disease (AD) patients. The total MMSE score is valuable for documenting cognitive changes over time, with AD patients typically experiencing an annual decline of 3 points on the MMSE (Bernard, Goldman, 2010). However, it's important to emphasize that the MMSE should not be solely relied upon for diagnosing dementia (Arevalo-Rodriguez, Smailagic, Roqué-Figuls, Ciapponi, Sanchez-Perez, Giannakou *et al.*, 2021). This caution arises from the potential influence of non-neurological factors on low scores, including factors like low education, language difficulties, and visual or auditory impairments. Other problematic aspects and limitations of the MMSE have been discussed in studies by Mitchell (2009), Tsoi (2015), and De Roeck (2019). In addition to its routine use in clinical practice, the Mini-Mental State Examination (MMSE) serves various other purposes. It is employed as an exclusion or inclusion criterion in clinical trials, primarily for screening cognitive impairment, and is included in the neuropsychological test batteries used in research studies. In Italy, the MMSE is a prerequisite for the accreditation of Nursing Homes under the National Health Service. It is typically administered to each patient within 30 days of admission and then repeated every 6 months according to the Individual Care Plan, except in cases of sudden changes that necessitate a new evaluation. Each Italian Region implements the national law with its specific Regional Council resolutions, such as DGR 7435/2001 and DGR 1765/2014 in Lombardy. Furthermore, the MMSE score is employed to categorize the severity of Alzheimer's disease (AD). For example, it is used to define a scale of severity in note n. 85 by the Italian Medicines Agency (AIFA)[1], which authorizes the prescription of AChE inhibitors (donepezil, rivastigmine, and galantamine) for mild and moderate AD, as well as memantine for moderate and severe AD. This note also outlines a staging

[1] https://www.aifa.gov.it/documents/20142/1728086/nota-85.pdf.

system based on MMSE scores: AD mild (MMSE 21-26), moderate (MMSE 10-20), moderately severe (MMSE 10-14), and severe (MMSE <10). This same scale is also adopted by the National Institute for Health and Clinical Excellence (NICE)[2] in England.

2.3 Analysis by means of linguistic features

Various methods can be employed to establish correlations between the linguistic production of speakers with cognitive impairments and clinical indexes like the MMSE. In several studies, linguistic features have been utilized for this purpose, resulting in correlation values that vary considerably. These variations depend on numerous factors, with the type of task or linguistic production playing a pivotal role. A comprehensive overview of these features and their presumed discriminatory power in dementia research can be found in de la Fuente Garcia *et al.* (2020: 1552) and Petti *et al.* (2020). Notably, the ADReSS Challenge at INTERSPEECH 2020, as described by Luz, Haider, de la Fuente Garcia, Fromm & Macwhinney (2020), defined two cognitive assessment tasks. The first task involved standard Alzheimer's speech classification, while the second task focused on neuropsychological score regression. In the latter task, participants were tasked with developing models to predict MMSE scores using audio recordings (and annotated transcriptions) obtained from participants who described the Cookie Theft picture (Luz *et al.*, 2021). A combination of acoustic and linguistic features directly extracted from audio recordings resulted in a baseline MMSE prediction root mean squared error (RMSE) of 5.28 (Luz *et al.*, 2020). A subset of participants reported and discussed their findings from the MMSE regression task, reporting a RMSE score between 4.54 (Balagopalan, Eyre, Robin, Rudzicz & Novikova, 2020) and 5.62 (Shah, Sawalha, Tasnim, Qi, Stroulia & Greiner, 2021). In Yeung, Iaboni, Rochon, *et al.*'s study (2021), a collection of features derived from Natural Language Processing (NLP) and Automated Speech Analysis (ASA) was extracted from 30 speech samples provided by individuals with Alzheimer's Disease (AD, MMSE: 15-20), those with MCI (MMSE: 23-26), and healthy controls (HCs, MMSE: 27-30), all having a minimum of 12 years of education. These same speech samples were evaluated by five clinicians, who rated them in terms of (1) word-finding difficulty, (2) incoherence, (3) preservation, and (4) errors in speech using a Likert scale (0=not present or normal finding; 1=mild; 2=moderate; 3=severe). The authors discovered statistically significant correlations between certain features and the clinical ratings. Similarly, Bueno-Cayo, Del Rio Carmona, Castell-Enguix, Iborra-Marmolejo, Murphy, Irigaray, Cervera & Moret-Tatay (2022) explored the relationship between MMSE scores and specific linguistic features in a cohort of 33 participants, consisting of 20 healthy controls and 13 individuals with Mild Cognitive Impairment (aged between 60 and 95, with basic or intermediate

[2] https://www.nice.org.uk/guidance/ta217/documents/alzheimers-disease-mild-to-moderate-donepezil-galantamine-rivastigmine-and-memantine-part-review-draft-scope2.

education). Their analysis revealed a positive correlation between MMSE scores and lexical density, which was measured as the ratio of semantic content words per 100 words in a sentence (Pearson's correlation = 0.488, $p<0.01$). Additionally, MMSE scores exhibited a negative correlation with age. However, no statistically significant correlation (α =0.05) was observed between MMSE scores and speech length, word frequency, or the number of tokens related to time, place, and action. In a multivariate linear regression analysis of MMSE scores, both lexical density and speech length emerged as significant predictors. The speech elicitation process involved a single open-ended question: "what are your plans for today and what are your plans for tomorrow," without any constraints on response time. However, in the study by Kavé, Dassa (2017), they observed negative correlations between MMSE scores and certain linguistic factors. Specifically, they found that MMSE scores had a negative correlation (Pearson) with the total number of words (-0.355, $p<0.5$) and with the mean word frequency (-0.339, $p<0.5$), indicating that greater dementia severity was associated with the production of more frequent words. On the other hand, MMSE scores were positively correlated with the types-to-token ratio (TTR) (0.572, $p<0.01$). These findings were based on a sample of 35 individuals with Alzheimer's disease (23 females, 12 males; age: 65-91; MMSE: 3-25; education: 8-16) who participated in a Cookie Theft picture description task. No statistically significant correlations were observed between MMSE scores and the content-word ratio, pronouns ratio, percentages of verbs, prepositions, and subordination markers. In contrast, Ostrand, Gunstad (2021) reported different results in their investigation. In the picture description task, they found positive correlations between the use of definite articles, determiners, and nouns as a percentage of the total word count and the Modified-Mini-Mental State Examination (3MS, Teng, Chui, 1987) scores. Additionally, they observed a negative correlation between the average lexical density and 3MS scores. Interestingly, in an expository speech task performed by the same individuals, only the average lexical density was found to be correlated with 3MS scores. The examples discussed here provide a glimpse into a complex landscape where different linguistic features can exhibit varying degrees of correlation with clinical indexes. These correlations are contingent upon several factors, with the type of task or linguistic production emerging as a pivotal determinant. For instance, in the study by Ostrand, Gunstad (2021), only lexical density showed a consistent correlation with 3MS scores across different tasks. Notably, this correlation was negative, contrasting with the positive correlation found in the study by Bueno-Cayo *et al.* (2022). It's important to highlight that, thus far, no strong correlation has been consistently observed between cognitive decline and linguistic and communication impairments (LCIs). Studies that have attempted to correlate linguistic characteristics with MMSE scores generally do not report a strong correlation. Some noteworthy exceptions do exist, primarily in works that focus on subjects with Alzheimer's disease (AD), such as the study by Fritsch, Wankerl & Nöth (2019), as opposed to more generalized cognitive impairment. In many studies aimed at identifying AD

or MCI, the approach involves categorization tasks and statistical comparisons of linguistic features between groups of unhealthy individuals (either AD or MCI) and matched healthy controls (HCs), as demonstrated in the work by Beltrami *et al.* (2018). Even when datasets encompass three or more diverse cohorts, most studies tend to limit their analyses to pairwise comparisons, typically contrasting dementia with healthy aging (Gagliardi, 2023). However, when considering the practical application of these findings in clinical contexts, the prevailing approach of pairwise comparisons may not be suitable for developing effective tools for real-life situations. This focus on pairwise comparisons often overlooks the intricate nature of assessing cognitive frailty within geriatric settings (Panza, Solfrizzi, Barulli, Santamato, Seripa, Pilotto & Logroscino, 2015). Some studies have exclusively considered or included comparisons between different "grades" of cognitive impairments, which aligns more closely with the approach presented in this work. For example, König, Satt, Sorin, Hoory, Toledo-Ronen & Derreumaux (2015) examined Alzheimer's disease (AD), MCI, and healthy controls (HCs) using vocal tasks like counting backward, sentence repetition, picture description, and verbal fluency. They identified significant differences ($p<0.05$) between AD and MCI only in the neuropsychiatric inventory (Cummings, Mega, Gray, Rosemberg-Thompson & Gornbein 1994). Only a limited number of research papers have delved into the detailed classification of dementia, with particular emphasis on subtyping frontotemporal degeneration (e.g., Fraser, Meltzer, Graham, Leonard, Hirst, Black & Rochon, 2014; Garrard, Rentoumi, Gesierich, Miller & Gorno-Tempini, 2014; Nevler, Ash, Irwin, Liberman & Grossman, 2019; Themistocleous, Eckerström & Kokkinakis, 2018, 2021; Cho, Nevler, Shellikeri, Ash, Liberman & Grossman, 2020). Regarding the classification techniques employed, numerous contributions have explored various machine learning algorithms, such as Support Vector Machines, Random Forests, and Decision Trees. These algorithms are used to automatically distinguish between healthy controls (HC) and individuals with MCI based on features extracted from acoustic and manually or automatically transcribed speech. These features encompass lexical, syntactic, semantic, and readability indexes, and researchers like Calzà *et al.* (2021) and Gagliardi, Tamburini (2021, 2022) have contributed to this area.

3. *Materials. The Anchise 2022 Corpus*

The Anchise Corpus, a valuable resource for research and training related to ILWD, has been continuously compiled since 2007 by the Anchise Group. This group comprises experts dedicated to the research, training, and care of ILWD. The corpus is founded on the principles of the "Enabling Approach," which was developed by one of the authors and collaborators back in 2004 (Vigorelli, 2004) and has evolved over subsequent years (Vigorelli, 2010, 2011; Lanzoni, Fabbo, Basso, Pedrazzini, Bortolomiol, Jones & Cauli, 2018; Vigorelli, 2021). The Enabling Approach is a non-pharmacological therapy for dementia that

aims to facilitate the speech of patients, even in the presence of memory and language disorders, with the goal of promoting a positive coexistence between individuals. To implement this approach, ILWD engage in conversations with trained operators from diverse backgrounds, including healthcare workers, educators, nurses, speech therapists, doctors, and psychologists, all of whom have undergone specific training in the Enabling Approach. According to the protocol, after initial greetings and obtaining informed consent, the operator initiates the conversation in a manner designed to encourage the patient to speak. This may involve phrases like "Could you tell me something about how you are going to spend your day?" or "What do you do during the day?" The operator then mostly remains silent, waiting for the elderly person's verbal production. When there is an interruption, the operator takes the floor and continues the conversation based on what the patient has said. Importantly, the operator does not lead the conversation but instead follows the patient's lead in their speech and world. The operator's role is not to diagnose or gather information but to participate in the conversation in a way that promotes the well-being of both speakers. Key techniques employed include refraining from asking questions, avoiding interruptions, refraining from corrections, echoing the patient's statements, and returning to the narrative motif. Other techniques involve listening attentively even when the speech is pathological and incomprehensible, respecting slowness and pauses, acknowledging and reflecting emotions, responding to questions, and communicating through gestures and tone of voice (Vigorelli, 2004, 2011). Conversations typically last 5-10 minutes, with no fixed duration, and conclude when signs of patient fatigue or intolerance become evident. It is important to note that since operators receive instructions in a training context, their adherence to the prompts may vary, unlike in a strictly controlled experimental setting. Since 2007, healthcare professionals affiliated with the Anchise Group have recorded, transcribed, and annotated these conversations, involving elderly individuals with cognitive impairment, primarily residing in Italian Nursing Homes. Specifically, the participants are Italian speakers with noticeable cognitive deficits, with inclusion criteria based on MMSE scores ranging from 0 to 28. Exclusion criteria encompass psychological or behavioural disturbances that hinder conversation, such as drowsiness, wandering, marked oppositional or aggressive behaviour, pronounced psychomotor agitation, severe dysarthria, and severe hypoacusia. The recordings are made with the prior informed consent of the relevant family member, the manager of the facility and, as far as possible, the elderly person participating in the conversation. The instrument is usually the operator's personal cell phone. The transcription is done by the conversant. A faithful transcription is required, including truncated words, neologisms, repetitions. The context annotations are freely indicated by the operator, someone does more, someone else does less. All transcriptions are incorporated into the Anchise Corpus and are intended for use in training sessions for Nursing Home operators, not for research purposes, without selecting them based on

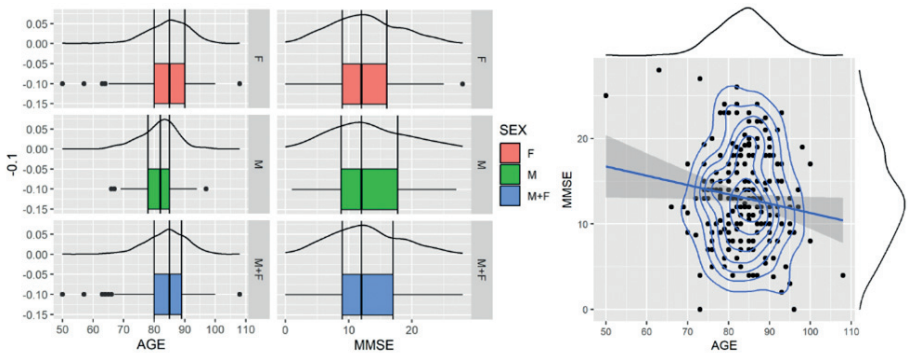
adherence to a specific style. For the same reason, the fidelity of the transcription is entrusted to the person sending the text and is not controlled. The corpus is continuously updated through the efforts of the Anchise Group. A prior version, the "Corpus Anchise 320," was released in 2020, comprising 320 conversations (Bolioli, Benvenuti, Mazzei & Vigorelli, 2020; Benvenuti, Bolioli, Bosca, Mazzei & Vigorelli, 2021). The Anchise 2022 Corpus extends this collection with 417 conversations. The significance of studying the Anchise Corpus lies in its representation of an Italian-language corpus collected in a natural environment. It serves as a valuable and rich source of information regarding the speech of individuals affected by dementia. The corpus reflects the diverse landscape found across Italy, including different regions (North, Center, and South), Nursing Homes of varying sizes, and patients of both sexes, with mild, moderate, and severe cognitive deficits. Patient classification is solely based on MMSE scores. Unlike prospective, controlled studies conducted under standardized conditions, this corpus was not originally created for research purposes but was intended for personnel training. Each dialogue or conversation within the corpus is organized into speech turns, with a new turn initiated each time there is a change in the speaker, continuing until the speaker talks without significant interruption. It is assumed that each conversation involves a different patient. Out of the 417 conversations, two were conducted in Spanish and English and were therefore excluded from the analysis. Some of the conversations include additional patient information, such as their gender, age, and corresponding MMSE score. For the experiments conducted, records were specifically selected based on patients with MMSE scores ranging from 0 to 26 (as detailed in § 4.1). These records were also required to include age information, providing more control over the subset of the corpus used in the experiments. This resulted in a total of 216 conversations. Table 1 provides initial insights into both the entire dataset and the filtered dataset. The interviewed patients have an average age of approximately 84 years, with a similar distribution between the entire corpus and the filtered records. The overall corpus is predominantly composed of female speakers, accounting for nearly 80 % of all speakers, and a similar gender ratio is observed in the filtered corpus. The conversations consist of an average of 65 turns, with a nearly even split between patient and operator turns. On average, turns contain over 14 tokens, although this figure significantly varies when considering patient turns (which tend to be longer, averaging around 20 tokens per turn) and operator turns (which are shorter, averaging about 9 tokens per turn). It is important to note that there is a high standard deviation, which is comparable to, if not greater than, the average itself, indicating substantial variation in turn length.

Table 1 - *Summary description of the Anchise 2022 Corpus composition with respect to subjects' type (Patients, Operators), patients' sex, speech turns, availability of MMSE score and age data*

	Corpus	*Filtered Corpus*
N. of conversations	417	216
Average patient age (stdev)	83.87 (7.09)	84.01 (6.93)
Male patients (percentage)	87 (21 %)	49 (23 %)
Female patients (percentage)	330 (79 %)	167 (77 %)
Average patient MMSE (stdev)	12.69 (5.60)	12.92 (5.46)
Average n. of turns per conversation	65.3	64.90
Average n. of patient turns per conversation	32.3	32.04
Average n. of operator turns per conversation	33	32.86
Average turn length (stdev)	14.25 (10.12)	14.83 (12.05)
Average turn length Patient (stdev)	19.60 (19.43)	20.31 (23.31)
Average turn length Operator (stdev)	8.85 (3.58)	9.24 (3.69)
Average conversation length (stdev)	831.68 (632.60)	843.50 (535.21)
Average conversation length – Patient (stdev)	543.76 (456.02)	545.82 (395.64)
Average conversation length – Operator (stdev)	287.92 (236.32)	297.67 (214.42)

Conversations in the filtered corpus are slightly longer than those in the broader corpus (averaging 843 tokens compared to 831 tokens). Again, this average length is accompanied by a notable standard deviation, suggesting consistent variations in length. Patient conversations are nearly twice as long as operator conversations. As depicted in Figure 1, their joint distribution of MMSE and age follows an approximately Gaussian distribution (Henze-Zirkler test = 0.86, p-value=0.15). There is a weak negative correlation between MMSE scores and age (β=-0.11, 95 % CI=[-0.21, -0.00426], t(217)=-2.05, p=0.041).

Figure 1 - *Distribution of the sessions with respect to Sex, Age, and MMSE scores. Left: boxplots with kernel density estimation (k.d.e.) for Age and MMSE by female, male and both. Right: scatterplot (dots), bivariate contour lines, linear trend and marginal k.d.e., for Age and MMSE (both male and female)*



4. *Method*

4.1 MMSE clustering and Evaluation Measures

In the original work, Folstein *et al.* (1975) reported that the MMSE should be scored from 0 to 30, with a score of 24 or greater as "normal" and with a score less than 20 "likely dementia." Afterwards, different classification scales (i.e. cut-off values) based on the MMSE score have been proposed or used in the literature. Some of these are illustrated as examples in Table 2 (see also Monroe, Carter, 2012, for further considerations). In this work only one severity scale was used, and the choice fell on the one recommended by AIFA, due to the national importance of the Agency, also supported by the adoption of the same scale abroad. However, the study could also be replicated for other scales. In the following, for sake of simplicity, the "AIFA85" term will be used to indicate the stages of cognitive impairments as depicted in the last row (*) of Table 2. The classification experiments considered only SEVERE, MODERATE and MILD stages, since the category of MMSE 27-30, has only 2 individuals. As can be seen by comparing (5) and (*) rows of Table 2, the "moderate-severe AD" (10-14) and "moderate AD" (15-20) grades were merged into a single "MODERATE" (10-20) grade. The reason for this merger lies primarily in the need to avoid complex classifications into four classes. Secondly, this grouping is also provided for by note 85 of AIFA, in the prescription of medications for patients with AD (memantine in severe patients with MMSE<=10; denoepezil, rivastigmine, galantamine, memantine, in moderate patients, with MMSE in 10 – 20, and donepezil, rivastigmine, galantamine in mild patients, with MMSE in 21 and 26).

Table 2 - *Some of the MMSE-based classification scales used in the literature. (1) De Stefano, Di Giovanni, Kulamarva, Di Fonzo, Massaro, Contini, Dispenza & Cazzato, 2021; (2) Tombaugh, McIntyre, 1992; (3): Chopra, Cavalieri & Libon, 2007; (4): Beltrami et al., 2018; (5): note n. 85 by the Italian Medicines Agency (AIFA), and National Institute for Health and Clinical Excellence (NICE), England. (*) = staging adopted in the present work, with number of individuals in parentheses. (mod.=moderate)*

<table>
<tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td><td>13</td><td>14</td><td>15</td><td>16</td><td>17</td><td>18</td><td>19</td><td>20</td><td>21</td><td>22</td><td>23</td><td>24</td><td>25</td><td>26</td><td>27</td><td>28</td><td>29</td><td>30</td><td></td></tr>
<tr><td colspan="10">Severe cogn. decline</td><td colspan="11">Mod. cogn. decline</td><td colspan="5">MCI</td><td colspan="5">Normal</td><td>(1)</td></tr>
<tr><td colspan="18">Severe cogn. impairment</td><td colspan="7">MCI</td><td colspan="5">Normal</td><td>(2)</td></tr>
<tr><td colspan="13">Severe impairment</td><td colspan="5">Mod. impair.</td><td colspan="7">Mild impair.</td><td colspan="3">Border-line</td><td colspan="3">Normal</td><td>(3)</td></tr>
<tr><td colspan="18"></td><td colspan="7">MCI or early dementia</td><td colspan="5">Normal</td><td>(4)</td></tr>
<tr><td colspan="9">Severe AD</td><td colspan="6">Mod.-severe AD</td><td colspan="8">Mod. AD</td><td colspan="6">Mild AD</td><td colspan="3"></td><td>(5)</td></tr>
<tr><td colspan="9">Severe (60)</td><td colspan="14">Mod. (137)</td><td colspan="6">Mild (19)</td><td colspan="3">Other (2)</td><td>(*)</td></tr>
</table>

Identifying correlations with MMSE scores involves developing evaluation tools capable of taking a transcript of a specific conversation as input and predicting the corresponding MMSE score for the patient participating in that conversation. Currently, we calculate

the Pearson and Spearman correlation coefficients (Benesty, Chen, Huang & Cohen, 2009) for this purpose. To evaluate the accuracy of the classification experiments, we used scores popular for Information Retrieval tasks, such as macro Precision, Recall and F1. In the multiclass classification task, for each class *c* we define:

- a *precision* score, as the number of samples of the class *c* that are correctly classified, divided by the total number of samples that are classified as *c*. The precision is the fraction of correct answers provided by the system.
- a *recall* score, as the number of samples of the class *c* that are correctly classified, divided by the number of input samples of the same class *c*. In other words, recall is the proportion of samples of class *c* that are correctly classified.
- a *F1* score, as the harmonic mean between precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The macro Precision is defined as the arithmetic mean of the precision across all the classes, and so macro Recall and macro F1.

4.2 Linguistic features

Leveraging the NLP tools of the "Stanza" software (Qi, Zhang, Y., Zhang, Y., Bolton & Manning, 2020), we extracted a range of morphosyntactic information and dependency relationships among the parts of speech from every turn within each patient session in the corpus (see Figure 2), and from these, also the linguistic features at the lexical, syntactic, and semantic levels shown in Table 3, for each session in the Corpus.

Figure 2 - *Top: a dependency tree among words of the sentence: Ecco... Io sono stato contento di essere qua con voi qua oggi ... ('well... I was happy to be here with you here today ...'). Bottom: a tabular representation of the parse tree, equipped with part-of-speech analysis obtained through the Stanza parser for the same sentence, in the CONLL-U format (Nivre, De Marneffe, Ginter, et al., 2016)*



id	form	lemma	upos	xpos	head	deprel	feats
1	Ecco	ecco	INTJ	I	6	discourse	
2	...	...	PUNCT	FS	1	punct	
3	Io	io	PRON	PE	6	nsubj	Number=Sing|Person=1|PronType=Prs
4	sono	essere	AUX	VA	6	aux	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
5	stato	essere	AUX	V	6	cop	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
6	contento	contento	ADJ	A	0	root	Gender=Masc|Number=Sing
7	di	di	ADP	E	9	mark	
8	essere	essere	AUX	V	9	cop	VerbForm=Inf
9	qua	qua	ADV	B	6	advmod	
10	,	,	PUNCT	FF	6	punct	
11	con	con	ADP	E	12	case	
12	voi	voi	PRON	PE	6	obl	Number=Plur|Person=2|PronType=Prs
13	qua	qua	ADV	B	14	advmod	
14	oggi	oggi	ADV	B	12	advmod	
15	...	...	PUNCT	FS	6	punct	

Table 3 - *Linguistic features*

Feature name (Lexical)	*Description*
content_density	the ratio between open-class words and open + closed class words (Roark, Mitchell, Hosom, Hollingshead & Kaye, 2011)
open_to_close_class	the ratio of open-class words to closed-class words
tokens_to_turns	the ratio between tokens and speech turns
num_words	the number of words
num_interrog	the number of interrogative sentences
num_ellipsis	the number of ellipsis «...»
word_length	the average number of characters of the words[3]
(Semantic & Lexical richness)	
num_hapax_lego	the number of words that appear only once in the session
num_types	the number of word types (i.e. different words)
ttr	types to token ratio
brunet indices	Brunet indexes. A measure of lexical diversity, used in stylometric analyses of text, often claimed to be independent of text length: $$BI = N^{V^{(-a)}}$$ where N is the text length, V is the number of different words, and −a is a scaling constant in a {0.134, 0.172, 0.185}
honore	Honoré statistics. An index of vocabulary richness, based on the idea that texts with richer vocabulary have a higher proportion of words that are hapax legomena. $$R = 100 \times \frac{\log(N)}{1 - \frac{v_1}{V}}$$ where N is the number of words v1 is the number of hapax, and V is the number of types
(Lexical / syntactic)	
<pos>_perc	Rate of: Adjectives (ADJ), Adpositions (ADP), Adverbs (ADV), Auxiliary (AUX), Coordinate conjunctions (CCONJ), Subordinate conjunctions (SCONJ), Nouns (NOUN), Pronouns (PRON), Interjections (INTJ), Proper nouns (PROPN), Verbs (VERB) E.g. adj_perc is the rate of Adjectives (ADJ) per 100 words.
sconj_to_cconj	ratio between subordinate and coordinate conjunctions (sentences)
nouns_to_verbs	nouns to verbs ratio. Also referred to as "reference rate to reality" (Vigorelli, 2004)
pronouns_to_nouns	the ratio between pronouns and nouns
x_perc	Rate of words that could match any other category (X) per 100 words.

[3] Even if there is no perfect correspondence between phone and characters, we considered this metric as it appears in other articles, such as in Orimaye, Wong, J.S., Golden, Wong, C.P. & Soyiri, 2017 and given that the Anchise corpus lacks audio tracks, which would be needed to count the number of the syllables and phones which have actually been realized.

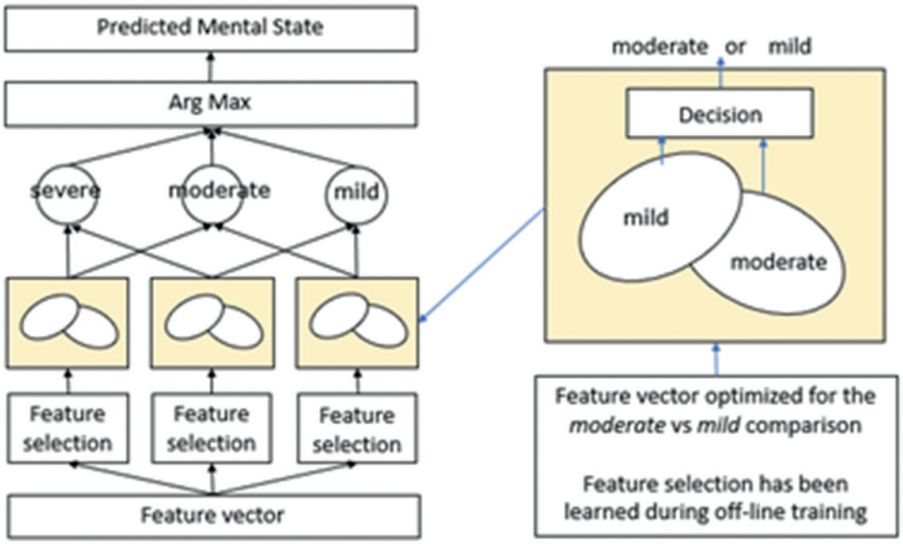
(Syntactic)	
dependency distance	Dependency distance (Roark, Mitchell & Hollingshead, 2007; Roark *et al.*, 2011). Mean, standard deviation and maximum values have been considered across each sentence.
max_depth	Maximum structure depth (mean, standard dev., maximum across each turn)
szmrecsanyi	A syntactic complexity measure by Szmrecsanyi, 2004: $\frac{2 * conj + 2 * pron + nouns + verbs}{words}$ Since subordinators and pronouns are considered as the clearest indicators of increased embeddedness (and thus of high complexity), these features have a higher weight than verbal forms and noun phrases.
(Verbal analysis)	
va_fin_rate, va_inf_rate, va_part_rate, va_ger_rate	Rates of finite, infinitive, participle, gerund verbal forms
reduced_sentences_count	n. of participle and gerund verbal forms
reduced_sentences_to_verbs	Ratio of participles and gerunds to total verbs
va_ind_rate, va_subj_rate, va_imp_rate, va_cond_rate	Rates of indicative, subjunctive(and conjunctive), imperative and conditional verbal moods
va_pres_rate, va_past_rate, va_imperf_rate, va_future_rate	Rates of present, past, imperfect and future tense
va_p1_rate	Rate of 1st person verbs

4.3 Correlation and classification between Linguistic Features and MMSE Scores

The features will be subject to analysis through linear regression, as well as Pearson and Spearman correlations, in relation to the MMSE scores.

Additionally, these same features will be employed for conducting statistical comparisons among different clusters of MMSE scores, categorized based on the AIFA85 stages. Due to variations in sample size across stages and the relatively small sample size in some cases, we have opted for the non-parametric two-sample Kolmogorov-Smirnov (KS) test. This test assesses the distance between the empirical distributions of individual features in both classes. We have established a significance level of α=0.05 as the threshold to determine statistical significance. Furthermore, we will conduct a subject classification task, which involves predicting the category of a session based on its linguistic features. This task is accomplished using the tool illustrated in Figure 3.

Figure 3 - *Scheme of the multiclass classifier used for the linguistic features classification experiment*



This tool consists of three binary classifiers, each designed to predict between two classes. The ultimate prediction for an input session is determined by the most frequently predicted output class. Each binary classifier is trained independently from the others, using only the sessions that belong to the two relevant classes. A leave-one-out cross-validation (LOOCV) approach is employed, where each session is taken out from the dataset and utilized for prediction, while the remaining sessions serve as the training dataset. We have tested two types of binary classifiers. The first assumes that the statistical distribution of the linguistic features for each of the two classes follows a Multivariate Normal distribution. During testing, it selects the class with the highest likelihood given the input session. The second classifier is based on Linear Discriminant Analysis (LDA), implemented using the Matlab® fitcd iscr function. During the training process, feature selection algorithms are applied to optimize performance metrics and identify a set of features that are especially responsive to the classification task. In this regard, we have explored Forward Feature Selection (FFsel) and Forward-Backward Feature Selection (FBsel). It is worth noting that the feature combination obtained through these methods may not necessarily represent the optimal choice, as there could be other combinations that yield equal or superior performance.

5. *Results*

5.1 Correlation between linguistic features and MMSE

Overall, the correlations between MMSE scores and individual linguistic features were relatively weak. However, some of these correlations did reach statistical significance ($p<0.05$). Table 4 displays these significant correlations, with the highest value being 0.19, representing the Pearson's correlation coefficient between MMSE scores and the frequency of verbs in the subjunctive mood. Regarding positive correlations, there are significant associations between MMSE scores and several linguistic features. Both Pearson and Spearman coefficients indicate a positive correlation between MMSE scores and the count of hapax legomena (i.e., words spoken only once in a session) as well as the rate of verbs in the subjunctive mood. Additionally, the rate of nouns, the rate of verbs in the past tense, the count of reduced sentences (i.e., the number of participle and gerund verbal forms), and the mean value of dependency distance are significantly correlated, though with only one of the correlation coefficients. Conversely, both Pearson and Spearman coefficients reveal significant negative correlations between MMSE scores and the rates of interjections, finite verbs, and the rate of verbs in the present tense. There is also a negative correlation, according to the Pearson coefficient, with the adjectives rate. Several features, including content density, showed no significant correlation with MMSE scores. Overall, while correlations between linguistic features and MMSE scores are weak, our findings align with existing literature.

Table 4 - *Linear fit (beta, value of the statistic, confidence interval, p-value and R-squared) and correlation values (Pearson, Spearman) between each single linguistic feature and MMSE scores. Only statistically significant results are reported (*p < 0.05, **p < 0.01, ***p < 0.001). Reduced sentence count is defined as the number of participle and gerund verbal forms. The label "(mean value of dependency distance" refers to the mean value of dependency distance. (r.o. = rate of)*

features	*beta*	*t*	*R2*	*Pearson*		*Spearman*	
				r	*p-value*	*ρ*	*p-value*
hapax count	0.0132	t(214)=2.250	0.023	0.152	0.025 *	0.145	0.033 *
r.o. adjectives	-0.4252	t(208)=-2.368	0.026	-0.162	0.019 *	-0.076	0.270
r.o. interj.	-0.1340	t(213)=-2.422	0.027	-0.164	0.016 *	-0.157	0.021 *
r.o. nouns	0.2124	t(212)=1.885	0.016	0.128	0.061 .	0.167	0.014 *
r.o. finite verbs	-0.0770	t(214)=-2.693	0.033	-0.181	0.008 **	-0.160	0.019 *
r.o. verbs in subjunctive mood	0.2622	t(214)=2.834	0.036	0.190	0.005 **	0.191	0.005 **
r.o.verbs in present tense	-0.0457	t(214)=-2.264	0.023	-0.153	0.025 *	-0.142	0.037 *
r.o. verbs in the past tense	0.0650	t(214)=2.054	0.019	0.139	0.041 *	0.120	0.078 .
n. reduced sentences	0.0374	t(214)=1.448	0.010	0.098	0.149	0.156	0.022 *
Dep dist mean avg	1.8419	t(214)=2.100	0.020	0.142	0.037 *	0.123	0.070 .

For example, we observed a positive correlation between the count of hapax legomena (words spoken only once) and MMSE scores, which is consistent with previous research indicating that individuals with greater cognitive impairment tend to have a more limited vocabulary. Additionally, our discovery of a positively correlated rate of nouns is in line with the concept of anomia, the initial linguistic disorder that often appears in individuals with cognitive impairments, particularly those with Alzheimer's disease (AD). Furthermore, the negative correlation we found between the rate of interjections and MMSE scores may suggest an increase in the level of uncertainty in speech production as MMSE scores decrease. These findings are consistent with previous research in the field. With respect to the mood and tense of verbal forms, as MMSE scores decline, we observe a trend toward simplification in the use of verbs. This trend aligns with real-world observations, indicating that individuals tend to employ more finite forms while reducing the use of participles, gerunds, and subjunctive moods. Moreover, there is a shift in the use of tenses, with a flattening effect towards the present tense at the expense of the past tense, as part of an overall simplification process. It's worth noting that the utilization of the future tense remains quite limited in these contexts. Several other features did not exhibit statistically significant correlations with MMSE scores. However, it's important to consider that certain speech-related or task-specific features may play a role. For instance, content density appears to be highly influenced by the nature of the speech task. To illustrate, a study by Beltrami, Calzà, Gagliardi, Ghidoni, Marcello, Rossini Favretti & Tamburini (2016) revealed that content density had statistical significance when comparing individuals with Mild Cognitive Impairment (MCI) to healthy controls in the context of a picture description task but not in a narrative task.

5.2 Classification

A statistical analysis was done on the distribution of each feature across the AIFA85 stages independently, utilizing the omnibus Kruskal-Wallis (KW) test and the Kolmogorov-Smirnov (KS) test. The KS test was applied even when the omnibus test did not yield significant results. Results are displayed in Table 5.

Table 5 - *Statistically significant differences between the distribution of the linguistic features in the AIFA85 stages. The statistics values and p-values of the Kruskal-Wallis (KW) and pairwise Kolmogorov-Smirnov tests are reported for each comparison (* p < 0.05, ** p < 0.01, *** p < 0.001). Bold values are significant after Bonferroni correction (p<0.05/3). (dep_dist. mean.avg = mean value of dependency distance)*

		SEVERE vs MODERATE d.o.f. (60, 137)		*SEVERE vs MILD d.o.f. (60,19)*		*MODERATE vs MILD d.o.f. (137,19)*	
	KW	*KS*	*p-value*	*KS*	*p-value*	*KS*	*p-value*
words count		0.167	0.165	0.384	0.019 *	0.257	0.182
hapax count	6.29 *	0.180	0.106	*0.428*	*0.006* **	0.334	0.035 *

		SEVERE vs MODERATE d.o.f. (60, 137)		*SEVERE vs MILD d.o.f. (60,19)*		*MODERATE vs MILD d.o.f. (137,19)*	
	KW	*KS*	*p-value*	*KS*	*p-value*	*KS*	*p-value*
average n. of words per turn		0.135	0.392	0.348	0.045 *	0.230	0.290
word-types count		0.130	0.419	*0.395*	*0.015* *	0.290	0.094 .
nouns to verbs ratio	6.32 *	*0.256*	*0.007* **	0.298	0.121	0.273	0.134
adpositions rate		0.118	0.550	0.361	0.034 *	0.260	0.171
nouns rate	8.25 *	*0.261*	0.005 **	0.279	0.170	0.192	0.499
rate of verbs in the subjunctive mood	11.3 **	*0.234*	*0.011* *	*0.439*	*0.002* **	0.264	0.135
dep_dist.mean.avg		0.147	0.294	0.359	0.036 *	0.258	0.179

In comparing the distribution of each single feature (independently of any other) in the 3 pairs formed by the 3 groups SEVERE, MODERATE and MILD, it is appropriate to adopt the Bonferroni correction. Therefore, results are presented and discussed with respect to a significance level α=0.05 as well a corrected α=0.05/3 = 0.017. The results indicate that four features are significant at the Kruskal-Wallis test, namely: the hapax count, the nouns-to-verbs ratio, the nouns rate and the rate of verbs in the subjunctive mood. Among these, three features, specifically the nouns to verbs ratio, nouns rate, and the rate of verbs in the subjunctive mood, display statistically different distributions when comparing SEVERE and MODERATE stages. Additionally, only one feature, the count of hapax legomena, exhibits distinct distribution between the MODERATE and MILD categories, but not after the Bonferroni correction. In contrast, eight features (three, after the Bonferroni correction) exhibit significant differences in distribution between the extreme categories, SEVERE and MILD. When comparing these results with the correlation analysis, it becomes evident that more features related to vocabulary richness exhibit statistical significance when distinguishing between the SEVERE and MILD stages. Specifically, the number of word-types (i.e., different words) and the count of hapax legomena, become significant (at different significance levels). This implies that the decline in vocabulary richness becomes more pronounced in the more severe stage. Notably, the significance of the hapax count, when shifting from MILD to MODERATE, is weaker compared to the comparison between the extreme categories. In the classification task, all available features were utilized, irrespective of whether they had been identified as significant or not. This approach aimed to ensure that no potential beneficial interactions between significant and non-significant features were overlooked. The most effective variant of the classifier discussed in § 4.3 employed multivariate normal (MVN) modelling of the feature distribution and employed Forward Backward selection (FBsel) as a feature selection strategy. The resulting evaluation metrics are presented in Table 6, which includes macro F1 scores, precision, and recall

scores for both the 3-classes and each individual 2-classes classification task. Most studies in the literature typically tackle a binary classification problem (e.g., AD vs. healthy or AD vs. MCI). Therefore, when comparing our results, we primarily focus on this aspect. In this context, our findings reveal that the highest accuracy is achieved in the classification of SEVERE vs. MILD. The F1 score, while not identical, is reasonably close to the results presented by Calzà *et al.* (2021), which also include some comparisons with other findings from the literature. However, it's essential to note that a direct comparison may not be entirely appropriate since the study by Calzà *et al.* involved the classification between MCI and healthy controls (F1=0.7045), and their feature set was more extensive, encompassing acoustic features as well.

Table 6 - *Summary results of the classification of the linguistic features. Macro F1, precision and recall*

Classification task	*F1*	*Precision*	*Recall*
SEVERE vs. MODERATE vs. MILD	0,533	0.562	0.537
SEVERE vs. MODERATE	0.630	0.627	0.636
SEVERE vs. MILD	0,779	0,785	0,773
MODERATE vs. MILD	0.596	0.595	0.691

6. *Discussion*

Analysing the transcribed dialogues obtained in ecologically relevant settings posed several challenges, especially when detailed sociolinguistic and diagnostic information was lacking. Nonetheless, despite its unique characteristics, both in terms of its objectives and data collection methods, the Anchise Corpus provided valuable qualitative and quantitative insights into speech patterns in individuals with dementia. Overall, the results of the analysis are in line with those reported in the literature. In summary, we did observe correlations between linguistic features in oral production and MMSE scores, although these correlations were generally weak. These findings align with the mixed results reported in the literature, as discussed in § 2.3, and may be attributed in part to the specific task and speech style we investigated, consistent with the outcomes in Beltrami *et al.* (2016). Nevertheless, the significant correlations we identified are in accordance with existing literature. For example, the positive correlation between the count of hapax legomena and MMSE scores supports findings in Hernandez-Dominguez, Ratte, Sierra-Martinez, *et al.* (2018). In essence, the significant correlations imply that individuals with more pronounced cognitive impairment tend to exhibit a reduced vocabulary and an increased level of uncertainty in their speech production, as evidenced by a higher frequency of interjections. Additionally, they appear to simplify their use of verbs by relying more on finite modes and present tense forms. Regarding the task of distinguishing between different stages of cognitive impairment based on MMSE

scores, the tools utilizing lexical and morphosyntactic features delivered promising accuracy results. For instance, in the binary classification task of distinguishing SEVERE from MILD cases, the F1 score approached 0.78. In the broader 3-class categorization task, the macro-averaged F1 score reached 0.53. These results stand up well to those reported in the literature, even when additional information and features, such as audio signals, were incorporated into the classification. Notably, the literature often focuses on distinguishing impaired individuals from healthy controls, as seen in the study by Calzà *et al.* (2021).

7. *Conclusions and Future Work*

Our research aimed to present the new extended release of the Anchise corpus, and to gain a deeper understanding of the speech patterns of ILWD engaged in natural conversations. We explored how linguistic features derived from NLP of the transcribed turn of speech could be correlated to the level of cognitive impairment, as assessed by the MMSE score. To achieve this goal, we analysed over 200 transcribed conversations from the Anchise 2022 Corpus. Our research revealed that there is generally a weak correlation between individual speech features and MMSE scores when examined individually. However, these features proved to be reasonably effective in distinguishing ILWD within MMSE-based clustering systems like NICE or AIFA85, particularly when differentiating between severe and mild cognitive impairments. This outcome aligns with the fact that our NLP experiments and MMSE tests focus on partially different aspects, as MMSE assessments are not specific to language or speech capabilities. This is the first experiment on a large corpus of spontaneous dialogues in Italian, but at the same time it is only a starting point. Indeed, the analysis of spontaneous speech by ILWD using the Anchise Corpus resources will be further expanded. In future works, we plan to integrate various analytical approaches, including a broader set of linguistic features, as well as the use of automatic emotion analysis and classification techniques employing Deep Language Models. This more comprehensive range of analyses is expected to provide a more detailed understanding of linguistic and communication deficits and offer more precise insights into the progression of dementia in relation to MMSE scores, surpassing the results obtained in this study.

Acknowledgements

This work was partially funded by the Italian Ministry, D.M. 1062/2021(within the innovation topics – Actions IV.4 – scientific sector L-LIN/01).

Bibliography

Arevalo-Rodriguez, I., Smailagic, N., Roqué-Figuls, M., Ciapponi, A., Sanchez-Perez, E. & Giannakou, A. et al. (2021) Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). In *Cochrane Database of Systematic Reviews*, 2021(7), CD010783. doi: 10.1002/14651858.CD010783.pub3

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F. & Novikova, J. (2021). Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. In *Frontiers in aging neuroscience*, 13(2021): 635945. doi: 10.3389/fnagi.2021.635945

Banovic, S., Zunic, L.J. & Sinanovic, O. (2018). Communication Difficulties as a Result of Dementia. In *Mater. Sociomed*, 2018, 30, 221–224. doi: 10.5455/msm.2018.30.221-224

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J. & McGonigle, K.L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. In *Arch Neurol*, 1994; 51(6):585–94.

Beltrami, d., calzà, l., gagliardi, g., ghidoni, e., marcello n., rossini favretti, r. & tamburini, f. (2016). Automatic identification of Mild Cognitive Impairment through the analysis of Italian spontaneous speech productions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2086–2093, Portorož, Slovenia. European Language Resources Association (ELRA).

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F. & Calzà, L. (2018). Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? In *Front. Aging Neurosci*. 2018, 10, 369. doi: 10.3389/fnagi.2018.00369

Benesty, J., Chen, J., Huang, Y. & Cohen, I. (2009). Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-00296-0

Benvenuti, N., Bolioli, A., Bosca, A., Mazzei, A. & Vigorelli, P. (2021). The "Corpus Anchise 320" and the analysis of conversations between healthcare workers and people with dementia. In Dell'Orletta, F., Monti, J. & Tamburini, F. (a cura di), *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*: Bologna, Italy, March 1-3, 2021. Torino: Accademia University Press. doi:10.4000/books.aaccademia.8260, 51-57.

Bernard, B., Goldman, J.G. (2010). MMSE – Mini-Mental State Examination. In Kompoliti K., Verhagen Metman, L. (a cura di), *Encyclopedia of Movement Disorders*, Academic Press, 2010, 187-189, doi: 10.1016/B978-0-12-374105-9.00186-6

Bolioli, A., Benvenuti, N., Mazzei, A. & Vigorelli, P. (2020). Analisi linguistica computazionale del "Corpus Anchise" di dialoghi operatore-paziente. In *Atti del 65° Congresso Nazionale Società Italiana di Gerontologia e Geriatria SIGG 2020*, Italy, 2-4 December, 2020.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A. & Cappa, S.F. (2017). Connected speech in neurodegenerative language disorders: a review. In *Front Psychol*. 2017; 8:269. doi:10.3389/fpsyg.2017. 00269

Bueno-Cayo, A.M. del Rio Carmona, M., Castell-Enguix, R., Iborra-Marmolejo, I., Murphy, M., Irigaray, T.Q., Cervera, J.F. & Moret-Tatay,

C. (2022). Predicting Scores on the Mini-Mental State Examination (MMSE) from Spontaneous Speech, In *Speech. Behav. Sci.* 2022, 12, 339. doi: 10.3390/bs12090339

Calzà, L., Gagliardi, G., Rossini Favretti, R. & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying Mild Cognitive Impairment and dementia. In *Computer Speech & Language*, 65, 101-113. doi: 10.1016/j.csl.2020.101113

Cho, S., Nevler, N., Shellikeri, S., Ash, S., Liberman, M.Y. & Grossman, M. (2020). Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features. In *RaPID@LREC*.

Chopra, A., Cavalieri, T.A. & Libon D.J. (2007). Dementia screening tolls for the primary care physician. In *Clin Geriatr*. 2007;15(1):38–45.

Cummings, J.L., Mega, M.S., Gray, K., Rosemberg-Thompson, S. & Gornbein, T. (1994). The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 1994; 44:2308–14. doi: 10.1212/wnl.44.12.2308

de la Fuente Garcia, S., Ritchie, C.W. & Luz, S. (2020). Artificial Intelligence, speech, and language processing approaches to monitoring Alzheimer's Disease: A systematic review. In *Journal of Alzheimer's Disease*, 78(4), 1547–1574. doi: 10.3233/JAD-200888

De Roeck, E.E., De Deyn, P.P., Dierckx, E. & Engelborghs, S. (2019). Brief cognitive screening instruments for early detection of Alzheimer's disease: a systematic review. In *Alz Res Therapy*, 11, 21(2019). doi: 10.1186/s13195-019-0474-3

De Stefano, A., Di Giovanni, P., Kulamarva, G., Di Fonzo, F., Massaro, T., Contini, A., Dispenza, F. & Cazzato, C. (2021). Changes in Speech Range Profile Are Associated with Cognitive Impairment. In *Dement Neurocogn Disord*. 2021 Oct;20(4):89-98. doi: 10.12779/dnd.2021.20.4.89

Dovetto, F.M., Guida, A., Pagliaro, A.C., Guardasci, R., Raggio, L., Sorrentino, A. & Trillocco, S. (2022, Online 2021). Corpora di Italiano Parlato Patologico dell'età adulta e senile. In Cresti, E., Moneglia, M. (a cura di), *Corpora e Studi Linguistici*. Atti del LIV Congresso della Società di Linguistica Italiana. Milano: Officinaventuno, 165-177.

Espinoza-Cuadros, F., Garcia-Zamora, M.A., Torres-Boza, D., Ferrer-Riesgo, C.A., Montero-Benavides, A., Gonzalez-Moreira, E. & Hernandez-Gomez, L.A. (2014) A spoken language database for research on moderate cognitive impairment: Design and preliminary analysis. In *Advances in Speech and Language Technologies for Iberian Languages. Lecture Notes in Computer Science* (2014), vol 8854. Springer, Cham. doi: 10.1007/978-3-319-13623-3_23

Filiou, R.P., Bier, N., Slegers, A., Houzé, B., Belchior, P. & Brambati, S.M. (2020). Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. In *Aphasiology*, 34(6), 723–755. doi: 10.1080/02687038.2019.1608502

Folstein, M., Folstein, S. & McHugh, P. (1975). "Mini-mental state"—A practical method for grading the cognitive state of patients for the clinician. In *J. Psychiatric Res.*, vol. 12, pp. 189–198, 1975. doi: 10.1016/0022-3956(75)90026-6

Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E. & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. In *Cortex*, 55, 43–60.

Fritsch, J., Wankerl, S. & Nöth, E. (2019). Automatic diagnosis of Alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing*. IEEE; 2019, 5841–5. doi: 10.1109/ICASSP.2019.8682690

Gagliardi, G. (2023). Natural language processing techniques for studying language in pathological ageing: A scoping review. In *International Journal of Language & Communication Disorders*. doi: 10.1111/1460-6984.12870

Gagliardi, G., Tamburini, F. (2021). Linguistic biomarkers for the detection of Mild Cognitive Impairment. In *Lingue e Linguaggio*, 1/2021, 3-31, doi: 10.1418/101111

Gagliardi, G., Tamburini, F. (2022). The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5234–5242, Marseille, France. European Language Resources Association.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B. & Gorno-Tempini, M.L. (2014) Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. In *Cortex*, 55, 122–129.

Goodglass, H., Kaplan, E. & Weintraub, S. (1983). Boston Naming Test. Lea and Febiger.

Hernández-Domínguez, L., Ratte, S., Sierra-Martinez, G. et al. (2018). Computer based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. In *Alzheimer's Dement*, 2018; 10: 260–8. doi: 10.1016/j.dadm.2018.02.004

Karr, J.E., Graham, R.B., Hofer, S.M. & Muniz-Terrera, G. (2018). When does cognitive decline begin? A systematic review of change point studies on accelerated decline in cognitive and neurological outcomes preceding mild cognitive impairment, dementia, and death. In *Psychol. Aging*, 33(2), 195–218. doi: 10.1037/pag0000236

Kavé G., Dassa A. (2017). Severity of Alzheimer's disease and language features in picture descriptions In *Aphasiology*, doi: 10.1080/02687038.2017.1303441

Kim, B.S., Kim, Y.B. & Kim, H. (2019). Discourse measures to differentiate between mild cognitive impairment and healthy aging. In *Front. Aging Neurosci*. 11(221). doi: 10.3389/fnagi.2019.00221

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O. & Derreumaux, A. (2015) Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dement Diagnosis. In *Assess Dis Monit*. 2015;1:112–124.

Krein, L., Jeon, Y., Miller Amberber, A. & Fethney, J. (2019). The assessment of language and communication in dementia: A synthesis of evidence. In *The American Journal of Geriatric Psychiatry*, 27(4), 363–377. doi: 10.1016/j.jagp.2018.11.009

Lanzoni, A., Fabbo, A., Basso, D., Pedrazzini, P., Bortolomiol, E., Jones, M. & Cauli, O. (2018). Interventions aimed to increase independence and well-being in patients with Alzheimer's disease. Review of some interventions in the Italian context. In *Neurology, Psychiatry and Brain Research*. Volume 30, 2018, 137-143, ISSN 0941-9500. doi: 10.1016/j.npbr.2018.10.002

Luz, S., de la Fuente, S.D. & Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. In *Proceedings of the LREC 2018 Workshop "Resources*

and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)", Kokkinakis D, ed. ELRA, Paris.

Luz, s., Haider, F., de la Fuente, S., Fromm, D. & Macwhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. *Proceedings of Interspeech*, 2020, 2172-2176, doi: 10.21437/Interspeech.2020-2571

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2021a). Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proceedings of Interspeech 2021*, Brno, Czechia, 30 August – 3 September 2021, 3780–3784. doi:10.21437/Interspeech.2021-1220

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm D. & MacWhinney, B. (2021b). Alzheimer's dementia recognition through spontaneous speech. In *Frontiers in computer science*, 3(2021): 780169.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M. & Christensen, H. (2019). Dementia Detection Using Automatic Analysis of Conversations. In *Computer Speech Lang*. 53, 65–79. doi:10.1016/j.csl.2018.07.006

Mitchell, A.J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. In *Journal of psychiatric research,* 43, no. 4(2009): 411-431. doi: 10.1016/j.jpsychires.2008.04.014

Monroe T, Carter M. (2012). Using the Folstein Mini Mental State Exam (MMSE) to explore methodological issues in cognitive aging research. Eur J Ageing. 2012 Jun 15;9(3):265-274. doi: 10.1007/s10433-012-0234-8

Nevler, N., Ash, S., Irwin, D.J., Liberman, M. & Grossman, M. (2019) Validated automatic speech biomarkers in primary progressive aphasia. In *Annals of Clinical and Translational Neurology*, 6(1), 4–14.

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D. et al. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (1659-1666).

Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P. & Soyiri, I.N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinformatics, 18(1), 34. https://doi.org/10.1186/s12859-016-1456-0

Ostrand, R. & Gunstad, J. (2021). Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults. In *Journal of Geriatric Psychiatry and Neurology*. 2021;34(5):357-369. doi:10.1177/0891988720933358

Panza, F., Solfrizzi, V., Barulli, M.R., Santamato, A., Seripa, D., Pilotto, A. & Logroscino, G. (2015) Cognitive Frailty: A Systematic Review of Epidemiological and Neurobiological Evidence of an Age-Related Clinical Condition. In *Rejuvenation Res*. 2015 Oct;18(5):389-412. doi: 10.1089/rej.2014.1637

Petti, U., Baker, S. & Korhonen, A. (2020), A systematic literature review of automatic Alzheimer's disease detection from speech and language. In *Journal of the American Medical Informatics Association*, 27(11), 1784–1797. doi: 10.1093/jamia/ocaa174

Pope, C., Davis, B.H. (2011). Finding a balance: The Carolinas Conversation Collection. In *Corpus Linguistics and Linguistic Theory*, 7(1):143–161. doi: 10.1515/cllt.2011.007

Prins, R.S., Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. In *Aphasiology,* 18, 1075–1091. doi: 10.1080/02687030444000534

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020.

Roark, B., Mitchell, M. & Hollingshead, K. (2007). Syntactic complexity measures for detecting Mild Cognitive Impairment. In: Cohen, K.B., Demner-Fushman, D., Frieman, C., Hirschman, L. & Pestian, J. (a cura di), *Proceedings of the Workshop BioNLP 2007: Biological, translational, and clinical language processing. ACL – Association for Computational Linguistics*, 1–8.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. In *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2081–2089. doi.org/10.1109/TASL.2011.2112351

Shah, Z., Sawalha, J., Tasnim, M., Qi, S., Stroulia, E. & Greiner, R. (2021). Learning language and acoustic models for identifying Alzheimer's dementia from speech. In *Frontiers in Computer Science*, 3(2021): 624659. doi: 10.3389/fcomp.2021.624659

Soria Lopez, J.A., González, H.M. & Léger, G.C. (2019). Alzheimer's Disease. Handb Clin. Neurol. 2019, 167, 231–255. doi: 10.1016/B978-0-12-804766-8.00013-3

Szmrecśanyi, B.M. (2004). On operationalizing syntactic complexity. In Purnelle, G., Fairon, C. & Dister, A. (Eds.), *Proc. of the 7th International Conference on Textual Data Statistical Analysis*, Presses Universitaires de Louvain, 1031–1038.

Teng, E.L., Chui, H.C. (1987). The Modified Mini-Mental State (3ms) Examination. In *J Clin Psychiatry*. 1987;48(8):314-318.

Themistocleous, C., Eckerström, M. & Kokkinakis, D. (2018) Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural In *Networks. Front. Neurol*. 2018, 9, 1–10.

Themistocleous, C., Ficek, B., Webster, K., den Ouden, D-B., Hillis, A.E. & Tsapkini, K. (2021) Automatic Subtyping of Individuals with Primary Progressive Aphasia. In *Journal of Alzheimer's Disease*, 79(2021) 1185–1194.

Toledo, C.M., Aluísio, S.M., Santos, L.B., Brucki, S.M.D., Trés, E.S., Oliveira, M.O. et al. (2017). Analysis of Macrolinguistic Aspects of Narratives from Individuals with Alzheimer's Disease, Mild Cognitive Impairment, and No Cognitive Impairment. In *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 31–40. doi:10.1016/j.dadm.2017.08.005

Tombaugh, T.N., McIntyre, N.J. (1992). The mini-mental state examination: a comprehensive review [see comment]. In *J Am Geriatr Soc*., 40(9): 922–935.

Tsoi, K.K, Chan, J.Y., Hirai, H.W., Wong, S.Y. & Kwok, T.C. (2015). Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. In *JAMA Intern Med*. 2015 Sep;175(9):1450-8. doi: 10.1001/jamainternmed.2015.2152

Ulatowska, H.K., Allard, L., Donnell, A., Bristow, J., Haynes, S.M., Flower, A. et al. (1988). Discourse performance in subjects with dementia of the Alzheimer type. In H.A. Whitaker (a cura di), *Neuropsychological studies of nonfocal brain damage: Dementia and trauma*. New York: Springer. doi: 10.1007/978-1-4613-8751-0_4

Vigo, I., Coelho, L. & Reis, S. (2022). Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review. In *Bioengineering*, 2022, 9, 27. doi: 10.3390/bioengineering9010027

Vigorelli P. (ed.) (2004). *La conversazione possibile con il malato Alzheimer*. Milano: FrancoAngeli.

Vigorelli P. (2010). *The ABC Group for caregivers of persons living with dementia: self-help based on the Conversational and Enabling Approach. Non-pharmacological therapies in dementia*; 3: 271-286.

Vigorelli P. (2011). *L'approccio capacitante. Come prendersi cura degli anziani fragili e delle persone malate di Alzheimer*. Milano: FrancoAngeli.

Vigorelli P. (2021). *Dialoghi imperfetti. Per una comunicazione felice nella vita quotidiana e nel mondo Alzheimer*. Milano: FrancoAngeli.

Weiner, J., Schultz, T. (2016). Detection of intra-personal development of cognitive impairment from conversational speech. In *Speech Communication*, 12. ITG Symposium, Paderborn, Germany, 1-5.

Weiner, J. Frankenberg, C., Telaar, D., Wendelstein, B., Schröder, J. & Schultz, T. (2016). Towards Automatic Transcription of ILSE – an Interdisciplinary Longitudinal Study of Adult Development and Aging. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), 2016.

Yeung, A., Iaboni, A., Rochon, E. et al. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia In. *Alz Res Therapy,* 13, 109(2021). doi: 10.1186/s13195-021-00848-x

DOI: 10.17469/O2111AISV000013

SYRIA CIRA IMPARATO, MARIA IZZO, OLGA LIGUORI, ERNESTO ORSINO, DONATA SENSALE, TINA SANTARPIA, GIULIA PERFETTO, GIOVANNA GISON, GLORIA GAGLIARDI

Caratteristiche pragmatiche dell'eloquio nel Disturbo dello Spettro Autistico (ASD): uno studio pilota su un campione di bambini italofoni residenti in Campania

Pragmatic characteristics of speech in Autism Spectrum Disorder (ASD): a pilot study on Italian-speaking children living in Campania

Pragmatics is widely considered the most impaired area in Autism Spectrum Disorder (ASD), especially in terms of implicit conversation norms, inferential processes, and narrative abilities. This study tackles this issue by examining the pragmatic and conversational characteristics of semi-spontaneous speech in a cohort of 17 Italian-speaking school-age children with High-Functioning Autism, compared to a Control Group of 17 typically developing (TD) peers. We analyzed the conversational turns on morphosyntactic, pragmatic and fluency levels. The results revealed several differences between the two groups: in particular, children with ASD produced less syntactically complex and fragmented texts, with a reduced use of textual connectors. They also produced fewer filled pauses and contributed little to the exchange with the communicative partner compared to their peers. Taken together, these findings support the notion of pragmatic language impairments in ASD but also reveal significant interactions between morphosyntactic, pragmatic and fluency levels that deserve further study.

Keywords: Autism Spectrum Disorders, language, communication, pragmatics, narrative abilities.

1. *Introduzione*

I Disturbi dello Spettro Autistico (ASD – *Autism Spectrum Disorders*) sono un gruppo eterogeneo di disturbi del neurosviluppo che esordiscono nei primi anni di vita e sono caratterizzati sotto il profilo clinico dalla presenza di difficoltà nell'area socio-emotiva e comunicativa e di *pattern* di comportamento, interessi o attività ristretti, ripetitivi (APA, 2013). Le condizioni cliniche racchiuse sotto questo "termine ombrello" sono collocabili lungo un *continuum* di gravità e presentano confini sfumati (Vivanti, 2021). Agli estremi del *continuum* ci sono da un lato i soggetti con autismo "a basso funzionamento" (*Low-Functioning Autism* – LFA), ossia con disabilità intellettiva e sviluppo cognitivo inferiore alle attese; dall'altro i soggetti con autismo ad "alto funzionamento" (*High-Functioning Autism* – HFA),

cioè con un quoziente intellettivo (QI) almeno nella norma, che li rende quasi indistinguibili dai coetanei a sviluppo tipico.

L'eterogeneità che contraddistingue l'ASD sul piano eziopatogenetico e sintomatologico si riflette nella dimensione linguistica: a un estremo si collocano infatti i soggetti che non hanno mai acquisito il linguaggio verbale ('non-verbali') e, all'altro, coloro che, pur sviluppando competenze linguistiche fonetico-fonologiche, morfo-sintattiche e lessicali nella norma (Arciuli, Brock, 2014), manifestano compromissioni più o meno significative nella più generale competenza comunicativa.

Come osservato fin dai primi resoconti clinici sull'autismo, il dominio centrale delle difficoltà linguistico-comunicative nei soggetti con ASD è la pragmatica. Specifiche problematicità che interessano questa competenza persistono durante l'intero corso della vita e sono presenti anche tra gli individui ad alto funzionamento con abilità linguistiche simili a quelle dei coetanei a sviluppo tipico: essi presentano invero difficoltà nell'applicazione della conoscenza linguistica in contesto (Naigles, Tek, 2017).

In particolare, i soggetti con ASD mostrano una iniziativa comunicativa spontanea scarsa o del tutto assente e usano il linguaggio a scopo comunicativo in modo limitato, ossia raramente fanno commenti, chiedono informazioni o chiarimenti all'ascoltatore (Tager-Flusberg, Paul & Lord, 2005). Altre difficoltà si manifestano nella forma di violazioni delle regole implicite della conversazione (Vivanti, 2021): per esempio, faticano a mantenere la reciprocità e l'alternanza dei turni dialogici (Biancalani, Gagliardi & Innocenti, 2023), a portare avanti il *topic* proposto dall'interlocutore (Wilkinson, 1998), e spesso intervengono con osservazioni irrilevanti o estranee alla conversazione, mostrando una scarsa capacità nel giudicare quali informazioni siano realmente necessarie ai fini dello scambio comunicativo (Paul, Cohen, 1984). Inoltre, hanno difficoltà a usare e interpretare correttamente il linguaggio non-letterale (es. metafore, ironia, richieste indirette) e a compiere inferenze, cioè a cogliere i contenuti non asseriti esplicitamente nella comunicazione (Dennis, Lazenby & Lockyer, 2001; Young, Diehl, Morris, Hyman & Bennetto, 2005).

Anche le abilità narrative risultano spesso compromesse nella popolazione autistica: i dati ottenuti da compiti di *storytelling* rivelano, infatti, che i soggetti con ASD hanno difficoltà a rappresentare e raccontare agli altri il nucleo centrale (*gist*) di una storia, a causa dell'incapacità di riprodurre la catena significativa degli eventi che costituiscono una narrazione (Diehl, Bennetto & Young, 2006); si perdono spesso in contenuti inappropriati, inserendo enunciati o frammenti di enunciati ripresi da conversazioni precedenti o da altre fonti (es. cartoni animati, canzoni) che il più delle volte non hanno alcun legame semantico o pragmatico con il contesto comunicativo; faticano a considerare la prospettiva dell'ascoltatore presentando gli eventi in maniera confusa, disorganizzata e poco coesa (Pfanner, Tancredi & Marcheschi, 2008).

Sulla base di queste osservazioni, il presente studio si propone di individuare e descrivere le caratteristiche pragmatiche che contraddistinguono l'eloquio di una coorte di bambini italofoni in età scolare che hanno ricevuto diagnosi di Disturbo dello Spettro Autistico ad alto funzionamento, rilevando eventuali differenze significative rispetto a un gruppo di controllo formato da coetanei a sviluppo linguistico e cognitivo tipico. A questo scopo, si è scelto di valutare lo scambio conversazionale e il discorso narrativo, considerati i due terreni più ecologici per sondare l'uso del linguaggio sociale (Padovani, Pattuzzi, 2017).

Il progetto di ricerca è stato approvato dal Comitato di Bioetica dell'*Alma Mater Studiorum* – Università di Bologna (n. 0173455/2022). I *caregiver* di tutti i bambini coinvolti nello studio hanno espresso il loro consenso alla partecipazione del minore alla ricerca e al trattamento dei suoi dati personali.

2. *Metodologia*

2.1 Campione

Il campione reclutato per lo studio si compone di 34 bambini (26M, 8F) di età compresa fra i 6 e i 13 anni (9;9 ± 2;5), divisi in due gruppi bilanciati per sesso (test χ^2, *p-value* > 0.05), età (test di Wilcoxon-Mann-Whitney, *p-value* > 0.05) e provenienza geografica.

Il primo gruppo (ASD) è formato da 17 soggetti, di cui 13 maschi e 4 femmine, che hanno ricevuto diagnosi di Disturbo dello Spettro Autistico ad alto funzionamento, reclutati presso il Centro Medico Riabilitativo di Pompei (NA)[1]. Il secondo (GC – Gruppo di Controllo) è costituito da 17 pari di età (13M, 4F) con sviluppo cognitivo e linguistico tipico, rispetto ai quali verranno operati i confronti per le variabili linguistiche di interesse. I bambini sono stati coinvolti nel progetto per volere dei loro genitori, che si sono mostrati da subito interessati all'argomento e agli esiti della ricerca.

Al fine di contenere gli effetti della variazione diatopica, si è ritenuto più opportuno selezionare soltanto bambini residenti nella regione Campania, aventi l'italiano come lingua materna e il campano come comune sostrato dialettale.

2.2 Raccolta dati

I dati sono stati raccolti audioregistrando sessioni di parlato semi-spontaneo elicitato tramite la somministrazione, in forma di gioco, delle seguenti prove:

1. Descrizione di figura complessa: "*Cookie Theft*" (Goodglass, Barresi, Weintraub, Segal, Goodglass & Kaplan, 2001) (Fig. 1);
2. Generazione di storia ("*telling*"): "Una festa di compleanno" (tratta da "Sequenze da raccontare", Shubi, Fig. 2);

[1] La disparità relativa al genere è conforme con quanto descritto dal DSM-5 (APA, 2013), secondo cui l'ASD registra una prevalenza di 4.2 volte maggiore nei maschi rispetto alle femmine in età prescolare.

3. Rievocazione di storia ("*retelling*"): "*Bus Story Test*" (Renfrew, 2015; adattamento italiano a cura di Mozzanica *et al*., 2016) (Fig. 3)

Figura 1 - *Cookie Theft*



Figura 2 - *Una festa di compleanno*



Figura 3 - *Bus Story Test*



La raccolta dati è avvenuta tra giugno e novembre 2022, dopo aver ricevuto il consenso scritto da parte dei genitori di tutti i partecipanti allo studio. Le registrazioni sono state effettuate tramite cellulare in formato WAV (*Waveform Audio File Format*). I bambini del gruppo ASD sono stati osservati presso il centro di riabilitazione, ciascuno all'interno della stanza assegnata per il trattamento; mentre i bambini del GC sono stati osservati presso la propria abitazione oppure quella della sperimentatrice.

2.3 Analisi dei dati

Le registrazioni raccolte sono state trattate in ambiente ELAN (EUDICO Linguistic Annotator) (Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006) e quindi sottoposte a segmentazione, trascrizione ortografica e annotazione. La segmentazione è stata realizzata scegliendo come unità di allineamento l'enunciato, inteso come qualsiasi unità linguistica dotata di autonomia prosodica e pragmatica e delimitata nel flusso del discorso da un *break* prosodico percepito come terminale (Malvessi Mittmann, Barbosa, 2016; Moneglia, Cresti, 2006). Ciascuna unità prosodico-pragmatica riconosciuta è stata sottoposta a una trascrizione ortografica

in formato CHAT-Lablita – una versione del formato standard CHAT arricchito dell'annotazione del *parsing* prosodico (Cresti, Moneglia, 2005) – e, infine, annotata manualmente.

Nel complesso, sono state realizzate 102 trascrizioni (51 per gruppo) che, una volta convertite in formato .txt, sono state esaminate allo scopo di tracciare il profilo delle competenze morfosintattiche e pragmatiche dei bambini osservati.

In primo luogo, è stata esaminata la composizione morfosintattica e l'interpretabilità semantica delle unità individuate, distinguendo tra i) turni "vuoti" (silenzi), ii) turni contenenti soli fenomeni paralinguistici/extralinguistici (tosse, risata, ecc.), iii) turni costituiti esclusivamente da parole incomprensibili, iv) sequenze interrotte (per l'intervento di fattori esterni o cambi di pianificazione del discorso da parte del soggetto), v) enunciati sintatticamente non completi e/o semanticamente non interpretabili in maniera indipendente dal contesto, vi) vere e proprie frasi (interpretabili senza bisogno di un aggancio diretto alla situazione comunicativa e dotate, sotto il profilo sintattico, di un predicato, istanziato da un verbo di modo finito. Cfr. Sabatini, Camodeca & De Sanctis, 2014; Serianni, Della Valle, 2017), e vii) discorsi riportati (inclusi nella storia originale o inventati, riconoscibili grazie alla presenza di una cornice citante).

Nella categoria degli enunciati sintatticamente non completi e/o semanticamente non interpretabili in maniera indipendente dal contesto sono state inserite unità di vario tipo, tra cui frasi *verbless*, turni costituiti esclusivamente da esitazione, costruzioni frastiche la cui valenza verbale non risulta satura o caratterizzate da verbi usati in maniera inadeguata, sintagmi nominali che fungono da innesto a frasi subordinate relative.

Nella categoria delle frasi, infine, sono state incluse tutte le strutture i) dotate di piena completezza semantica/sintattica e ii) chiaramente riconoscibili nella loro valenza illocutiva, anche se interrotte sul piano prosodico. In generale, le costruzioni frastiche sono state analizzate guardando alla loro organizzazione interna e distinte, dunque, in principali, coordinate, subordinate e frasi incidentali. L'individuazione delle proposizioni coordinate e subordinate si è fondata sulla presenza di nessi espliciti (es. congiunzioni). Per frasi incidentali si considerano, in questa sede, soltanto quelle frasi che, inserite all'interno di altre frasi a esse non collegate sul piano sintattico, presentano un verbo – esprimente opinione, giudizio o dubbio (es. "penso", "credo") – collocato all'inizio, alla fine o nel mezzo della frase all'interno della quale è innestato. Anche le subordinate sono state ulteriormente indagate, andando a osservarne la tipologia e operando una distinzione tra argomentali, relative e avverbiali.

Sotto il profilo pragmatico, si è innanzitutto osservata la capacità di compiere inferenze, valutando, nei termini di presenza/assenza, se il soggetto fosse in grado di cogliere il collegamento implicito tra le immagini 3 e 5 contenute nella sequenza "Una festa di compleanno" (cfr. Fig. 2).

È stata poi analizzata la capacità dei soggetti di conferire coesione alle proprie produzioni, valutando nella fattispecie:

– tipologia e appropriatezza delle espressioni referenziali usate per riferirsi ai protagonisti e agli elementi cardine dell'immagine descritta o della storia raccontata, distinguendo tra sintagma nominale definito (SND), sintagma nominale indefinito (SNI), pronome tonico (Pron tonico), pronome atono (Pron atono), pronome relativo (Pron relativo), pronome nullo (Pron nullo), pronome dimostrativo (Pron dimostrativo) e sintagma nominale senza determinante (SNsD), e andando a verificare il modo in cui vengono gestite prima menzione e riprese successive;
– numero di connettivi testuali, includendo tutti quegli elementi in grado di reggere la relazione tra le frasi o le varie parti che costituiscono l'ossatura di una narrazione (per es. congiunzioni, preposizioni, avverbi, espressioni usate per disporre gli eventi in ordine logico e cronologico, ecc.);
– gestione e tipologia del discorso riportato.

In ultima istanza, si è provveduto a un'analisi delle produzioni orali sul piano della fluenza, valutando la durata del *task*, il numero e la durata di pause piene e pause vuote (silenzi) – i cui valori sono stati estratti automaticamente tramite ELAN – e, infine, il *Total Phonation Time* (TPT), indicativo del contributo verbale che il soggetto dà allo scambio comunicativo, rilevato sottraendo la durata totale delle pause vuote alla durata complessiva dei turni conversazionali ascrivibili al soggetto stesso. Nella Tab. 1 sono sintetizzati gli indici considerati.

Tabella 1 - *Indicatori considerati nello studio per ciascuno dei livelli indagati*

Livello morfosintattico
Totale delle unità (prosodico-pragmatiche) prodotte da ogni partecipante, considerando le tre prove nel complesso
Frequenza assoluta e percentuale di turni vuoti, turni con fenomeni paralinguistici/ extralinguistici, turni con sole parole incomprensibili, sequenze interrotte, enunciati sintatticamente non completi e/o semanticamente non interpretabili in maniera indipendente dal contesto, frasi, discorsi riportati
Livello pragmatico
Capacità di fare inferenze
Numero complessivo, tipologia e appropriatezza delle espressioni referenziali
Gestione del discorso riportato
Numero di connettivi testuali
Fluenza
Durata del *task*
Pause piene (numero e durata)
Pause vuote (numero e durata)
TPT

3. *Risultati*

I risultati ottenuti dalle analisi relative a ciascuno dei tre livelli indagati hanno rivelato significative differenze tra i due gruppi, sia dal punto di vista qualitativo che sotto il profilo statistico[2].

3.1 Livello morfosintattico

La Tab. 2 mostra i dati emersi dall'analisi della composizione dei turni conversazionali: nella prima colonna, vengono presentati gli indici indagati; nella seconda e terza colonna sono sintetizzate media e deviazione standard relative ai vari parametri estratti, rispettivamente, per il gruppo ASD e il GC. Nella quarta e quinta colonna, vengono mostrati *p-value* e significatività statistica di ciascuno degli 8 indici considerati, calcolati tramite il Test non-parametrico di Wilcoxon-Mann-Whitney. Le uniche differenze rilevate riguardano il totale delle unità prodotte e il numero di frasi realizzate in media dai due gruppi.

Nelle Figure 4 e 5 vengono inoltre illustrate, tramite due grafici a torta, le medie calcolate sui due gruppi relative alle diverse tipologie di unità rintracciate nei turni conversazionali, espresse in percentuale. I valori percentuali sono stati confrontati tramite il Test χ^2 con correzione di Yates, che ha restituito un *p-value* = 0.3843, indicativo di assenza di differenze rilevanti.

Tabella 2 - *Composizione dei turni (in parentesi, valori espressi in %)*

COMPOSIZIONE DEI TURNI				
	ASD	*GC*	*p-value*	*significatività*
Tot. Unità	100.53 ± 25.54	125.59 ± 27.94	0.009255	**
Turni vuoti	8.24 ± 10.49 (8.56 ± 11.64)	2.59 ± 2.53 (2.23 ± 2.44)	0.2502	/
Turni para-/ extralinguistici	1.65 ± 2.50 (1.41 ± 2.09)	0.65 ± 0.93 (0.46 ± 0.64)	0.4689	/
Turni con parole incomprensibili	0.35 ± 1.06 (0.24 ± 0.72)	0.00 ± 0.00 (0.00 ± 0.00)	0.1635	/
Sequenze interrotte	8.82 ± 4.46 (8.76 ± 4.39)	10.35 ± 4.37 (8.25 ± 3.25)	0.4369	/
Discorsi riportati	2.29 ± 1.61 (2.60 ± 2.10)	2.76 ± 0.97 (2.28 ± 0.93)	0.3024	/
Enunciati non completi sint./sem.	33.06 ± 19.43 (31.51 ± 11.62)	34.18 ± 11.80 (27.03 ± 6.96)	0.3344	/
Frasi	46.12 ± 19.34 (46.92 ± 19.89)	75.06 ± 19.72 (59.75 ± 8.33)	0.0004992	***

[2] La soglia di significatività per tutti i test statistici impiegati è stata fissata a *p-value* < 0.05. Nello specifico: p < 0.05 ("*"); p < 0.01 ("**"); p < 0.001 ("***"); p < 0.0001 ("****").

Figura 4 - *Composizione dei turni in percentuale; media calcolata sul gruppo ASD*

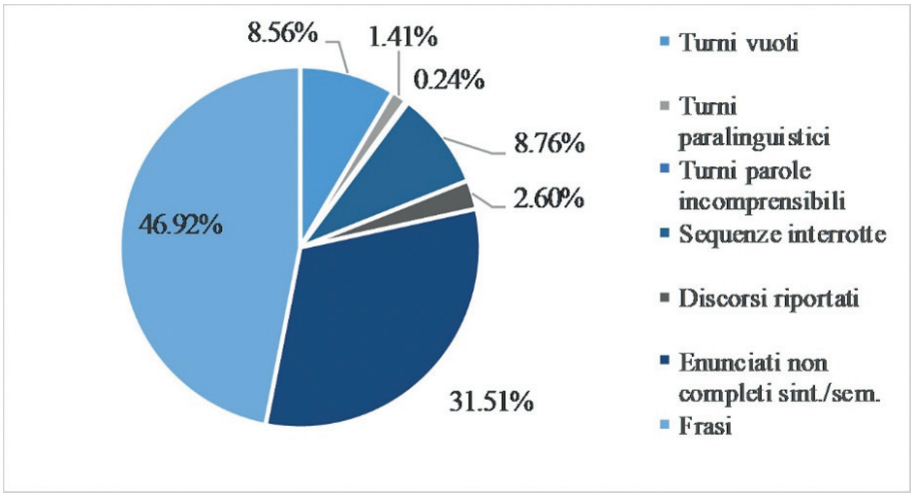
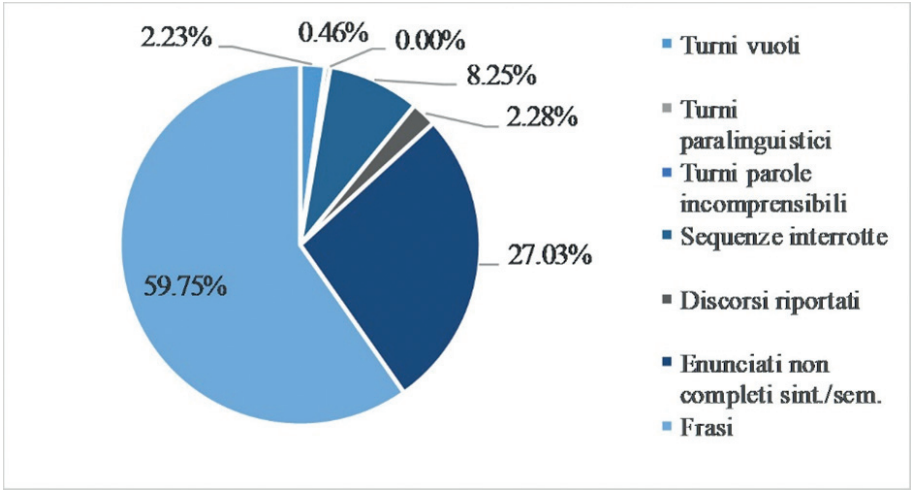


Figura 5 - *Composizione dei turni in percentuale; media calcolata sul GC*



Una seconda analisi ha indagato la composizione delle strutture frastiche prodotte dai due gruppi. I risultati ottenuti hanno segnalato differenze significative sia sotto il profilo qualitativo che quantitativo. Più nel dettaglio, la Tab. 3 presenta gli indici osservati (prima colonna), media e deviazione standard calcolate per i due gruppi (seconda e terza colonna), e, infine, *p-value* e significatività statistica dei parametri (quarta e quinta colonna), ottenuti applicando il Test di Wilcoxon-Mann-Whitney, che ha rilevato significatività statistica per ciascuno degli indici indagati, escluso il numero di principali.

Sono stati inoltre comparati i valori, espressi in percentuale, delle medie relative alle quattro diverse tipologie frasi prodotte dai due gruppi, attraverso l'applicazione

del Test $\chi 2$ con correzione di Yates, che ha restituito un *p-value* = 0.003051, indicativo della presenza di significatività statistica (**).

Tabella 3 - *Composizione delle strutture frastiche (in parentesi, valori espressi in %)*

COMPOSIZIONE DELLE STRUTTURE FRASTICHE				
Indice	*ASD*	*GC*	*p-value*	*significatività*
F_Principali	26.12 ± 9.73 (63.81 ± 24.19)	29.0 ± 9.50 (38.71 ± 7.58)	0.4081	/
F_Coordinate	10.53 ± 8.29 (18.90 ± 12.96)	19.65 ± 5.95 (26.32 ± 5.42)	0.002803	**
F_Subordinate	9.41 ± 7.31 (17.20 ± 12.14)	25.18 ± 8.79 (33.32 ± 7.91)	4.747e-05	****
F_Incidentali	0.06 ± 0.24 (0.09 ± 0.38)	1.18 ± 1.33 (1.64 ± 2.03)	0.002351	**

Una terza analisi ha esaminato le frasi subordinate impiegate dai bambini durante gli scambi comunicativi con l'intervistatore. Nella Tab. 4 sono sintetizzati gli indici osservati (prima colonna), media e deviazione standard relative a ciascun parametro estratto per ASD e GC (rispettivamente, seconda e terza colonna), e *p-value* e significatività statistica dei parametri (rispettivamente, quarta e quinta colonna), ottenuti attraverso l'applicazione del Test di Wilcoxon-Mann-Whitney. I risultati mostrano importanti disparità tra i due gruppi, dal momento che ciascuno degli indici osservati ha raggiunto significatività statistica.

Tabella 4 - *Tipologia delle subordinate (in parentesi, valori espressi in %)*

TIPOLOGIA DELLE SUBORDINATE				
Indice	*ASD*	*GC*	*p-value*	*Significatività*
S_ Argomentali	3.18 ± 3.03 (29.92 ± 27.58)	6.59 ± 2.79 (26.28 ± 7.97)	0.003198	**
S_ Relative	2.47 ± 2.70 (19.57 ± 19.13)	7.59 ± 4.06 (28.98 ± 12.88)	0.0002685	***
S_ Avverbiali	3.76 ± 3.35 (38.74 ± 31.61)	11.00 ± 4.66 (44.74 ± 14.05)	4.48e-05	****

3.2 Livello pragmatico

I dati ottenuti dall'analisi della capacità di cogliere l'inferenza contenuta nella sequenza di immagini "Una festa di compleanno" (cfr. Fig. 2) sono stati trasformati nella matrice di confusione riportata in Tab. 5, a cui è stato applicato il Test esatto di Fisher. Malgrado l'evidente differenza tra i due gruppi, segnalata dal numero più elevato di soggetti con ASD che non sono riusciti a cogliere l'implicito, sotto il profilo quantitativo è stato rintracciato un *p-value* = 0.08545.

Tabella 5 - *Capacità di cogliere inferenze: matrice di confusione*

	Y	*N*
ASD	11	6
CG	16	1

Rispetto agli indicatori di coesione testuale, la Tab. 6 presenta i risultati ottenuti dall'analisi della tipologia delle espressioni referenziali. Più nel dettaglio, la prima colonna riassume gli indici indagati; la seconda e terza colonna sintetizzano media e deviazione standard relativa a ciascun parametro estratto, rispettivamente, per il gruppo ASD e per il GC; la quarta e quinta colonna mostrano *p-value* e significatività statistica dei parametri, ottenuti tramite il Test di Wilcoxon-Mann-Whitney. Come è possibile osservare, soltanto i valori di Pron atono, Pron relativo e SNsD hanno raggiunto significatività statistica.

Tabella 6 - *Tipologia delle espressioni referenziali (in parentesi, valori espressi in %)*

TIPOLOGIA DELLE ESPRESSIONI REFERENZIALI				
Indice	*ASD*	*GC*	*p-value*	*Significatività*
Tot. Espressioni referenziali	51.47 ± 13.57	62.53 ± 12.24	0.06239	/
SNI	6.06 ± 3.51 (12.46 ± 9.76)	7.59 ± 2.53 (12.54 ± 4.80)	0.2053	/
SND	25.00 ± 7.49 (49.99 ± 13.37)	25.35 ± 8.27 (40.16 ± 8.05)	0.9587	/
Pron dimostrativo	0.53 ± 0.80 (0.88 ± 1.30)	0.47 ± 0.51 (0.76 ± 0.85)	0.8741	/
Pron tonico	1.18 ± 1.59 (1.92 ± 2.47)	2.29 ± 2.64 (3.80 ± 4.15)	0.1169	/
Pron atono	3.18 ± 2.90 (5.46 ± 4.70)	6.82 ± 4.16 (10.52 ± 5.09)	0.01676	*
Pron relativo	1.76 ± 2.08 (2.91 ± 3.27)	4.06 ± 2.82 (6.16 ± 3.29)	0,008091	**
Pron nullo	11.65 ± 6.78 (21.15 ± 11.77)	15.71 ± 3.79 (25.61 ± 6.88)	0.05706	/
SNsD	2.12 ± 4.51 (5.23 ± 11.20)	0.24 ± 0.44 (0.36 ± 0.68)	0.04619	*

La Tab. 7 mostra i risultati ottenuti dal confronto sul numero di connettivi testuali che il gruppo ASD e il GC hanno complessivamente utilizzato nel corso delle tre prove. La differenza tra i due gruppi è rilevante, poichè l'applicazione del Test di Wilcoxon-Mann-Whitney ha restituito un *p-value* = 0.003393.

Tabella 7 - *Numero di connettivi testuali*

ASD	*GC*	*p-value*	*significatività*
20.65 ± 17.67	38.29 ± 8.11	0.003393	**

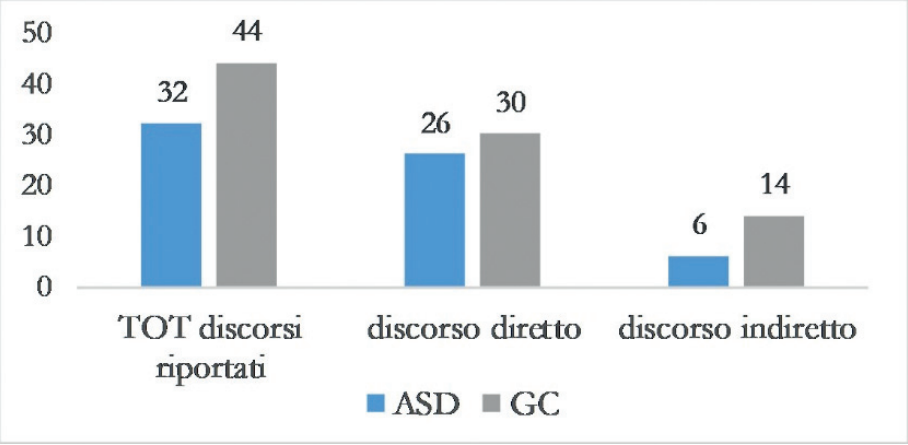
Accanto alle analisi quantitative fin qui descritte, è stata condotta un'analisi di carattere qualitativo sulla gestione del discorso riportato relativo al *Bus Story Test*.

La Tab. 8 mostra, nella prima riga, gli enunciati attesi dai due gruppi, che, tratti dal testo originale, sono stati pronunciati dall'intervistatore in una prima fase della somministrazione della prova, e, nella seconda e terza riga, rispettivamente, il numero di soggetti con ASD e il numero di controlli che hanno riportato in maniera più o meno fedele ciascuno degli enunciati considerati. Il confronto tra i due gruppi viene anche rappresentato graficamente nella Fig. 6, che illustra il totale dei discorsi riportati nel gruppo ASD e nel GC e il numero di discorsi diretti e indiretti rintracciati complessivamente in ciascuno dei due gruppi.

Tabella 8 - *Discorsi riportati*

	DISCORSI RIPORTATI			
	"Fermati autobus!"	*"Sono stufo di andare per la strada..."*	*"Muuuu!"*	*"Non credo ai miei occhi"*
ASD	10/17	6/17	6/17	10/17
GC	12/17	9/17	11/17	13/17

Figura 6 - *Gestione discorso riportato, ASD e GC a confronto*



3.3 Livello della fluenza

Le Tab. 9, 10 e 11 presentano i risultati parziali relativi agli indici di fluenza, rispettivamente, per "*Cookie Theft*", "Una festa di compleanno" e "*Bus Story Test*". In ciascuna tabella, sono riportati gli indici considerati (prima colonna), media e deviazione standard relativi ai parametri estratti per ASD (seconda colonna) e GC (terza colonna), e *p-value* e significatività statistica relativi a ciascun parametro (rispettivamente, quarta e quinta colonna), calcolati tramite il Test di Wilcoxon-Mann-Whitney. Le differenze significative tra i due gruppi riguardano, nella prova "*Cookie Theft*", la durata del *task* e il numero e la durata delle pause piene; nella prova "Una festa di compleanno", il TPT e il numero e la durata delle pause piene; nel "*Bus Story Test*", il TPT, il numero di pause vuote, e il numero e la durata delle pause piene.

Tabella 9 - *Fluenza "Cookie Theft"*

FLUENZA: COOKIE THEFT				
Indice	*ASD*	*GC*	*p-value*	*significatività*
Durata task	121.76 ± 78.64	62.47 ± 31.46	0.01378	*
Total Phonation Time *(TPT)*	45.88 ± 28.04	49.65 ± 17.52	0.2553	/
Pause vuote (numero)	22.65 ± 12.13	20.94 ± 6.14	0.863	/
Pause vuote (durata)	34.18 ± 31.41	17.47 ± 6.19	0.1568	/
Pause piene (numero)	2.18 ± 3.57	3.88 ± 3.22	0.03772	*
Pause piene (durata)	1.06 ± 2.41	2.12 ± 1.73	0.009818	**

Tabella 10 - *Fluenza "Una festa di compleanno"*

FLUENZA: UNA FESTA DI COMPLEANNO				
Indice	*ASD*	*GC*	*p-value*	*significatività*
Durata task	127.35 ± 94.72	74.47 ± 49.96	0.08189	/
Total Phonation Time *(TPT)*	55.94 ± 25.77	83.12 ± 38.68	0.0387	*
Pause vuote (numero)	26.12 ± 15.75	31.53 ± 14.31	0.1785	/
Pause vuote (durata)	36.18 ± 34.95	23.76 ± 11.93	0.8903	/
Pause piene (numero)	1.06 ± 1.43	6.00 ± 4.5	0.0001107	***
Pause piene (durata)	0.47 ± 0.72	2.59 ± 1.91	0.0001985	***

Tabella 11 - *Fluenza "Bus Story Test"*

FLUENZA: BUS STORY TEST				
Indice	*ASD*	*GC*	*p-value*	*significatività*
Durata task	128.41 ± 46.87	135.82 ± 37.49	0.4694	/
Total Phonation Time *(TPT)*	64.94 ± 19.08	88.12 ± 21.77	0.002035	**
Pause vuote (numero)	29.35 ± 8.32	37.94 ± 12.71	0.0455	*
Pause vuote (durata)	35.41 ± 19.78	26.71 ± 14.19	0.1733	/
Pause piene (numero)	0.82 ± 1.07	12.71 ± 8.96	5.804e-06	****
Pause piene (durata)	0.41 ± 1.00	5.88 ± 4.61	7.742e-06	****

Infine, la Tab. 12 mostra i risultati totali sul piano della fluenza, ottenuti dalla somma dei dati emersi dalle analisi condotte separatamente sulle tre prove (cfr. Tab. 9, 10 e 11). Il Test di Wilcoxon-Mann-Whitney ha rivelato significative differenze statistiche per il TPT e il numero e la durata delle pause piene. I *Box Plot* nelle Fig. 7, 8 e 9 offrono una rappresentazione visiva della variabilità dei valori emersi in totale per ciascuno dei due gruppi, evidenziando, quando serve, la presenza di *outlier*. Nella fattispecie, la Fig. 7 mostra il confronto tra GC e ASD in merito alla durata del *task* e al TPT, la Fig. 8 il numero e la durata di pause vuote, e, infine, la Fig. 9 il numero e la durata delle pause piene.

Tabella 12 - *Fluenza, risultati complessivi*

FLUENZA COMPLESSIVA				
Indice	*ASD*	*GC*	*p-value*	*significatività*
Durata task	377.53 ± 188.65	272.76 ± 94.68	0.06749	/
Total Phonation Time *(TPT)*	166.8 ± 47.49	220.88 ± 64.59	0.02198	*
Pause vuote (numero)	78.12 ± 26.62	90.41 ± 28.37	0.1735	/
Pause vuote (durata)	105.76 ± 79.58	67.94 ± 27.86	0.2485	/
Pause piene (numero)	4.06 ± 4.28	22.59 ± 13.79	7.175e-05	****
Pause piene (durata)	1.94 ± 2.79	10.59 ± 6.90	9.172e-05	****

Figura 7 - *Durata task totale (a sinistra) e TPT totale (a destra): GC e ASD a confronto*

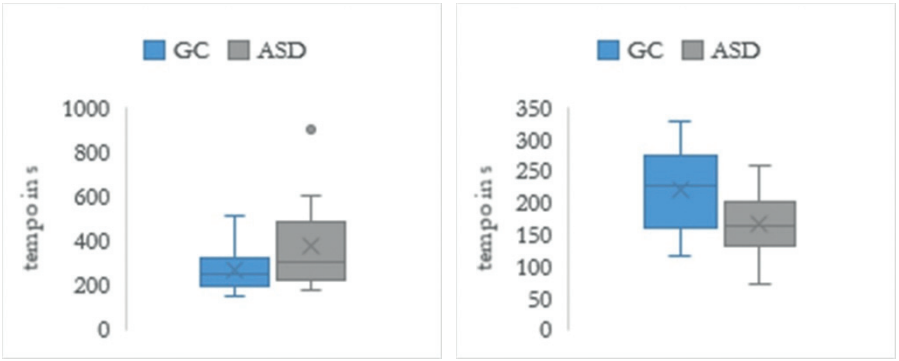


Figura 8 - *Numero pause vuote totale (sinistra) e durata pause vuote totale (destra): GC e ASD a confronto*

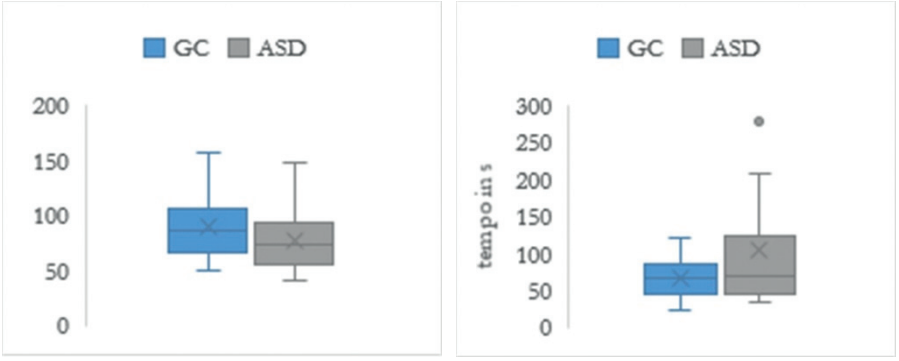
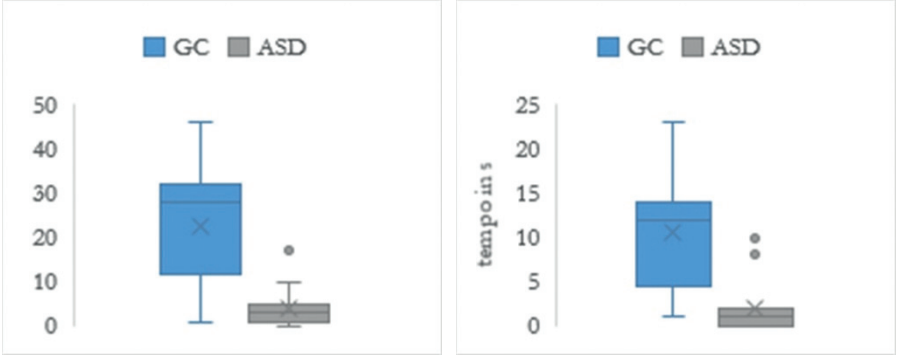


Figura 9 - *Numero pause piene totale (a sinistra) e durata pause piene (a destra): GC e ASD a confronto*



4. *Discussione*

Come dimostrato nel § 3, le varie analisi condotte hanno fatto emergere una serie di differenze sul piano linguistico e comunicativo tra i due gruppi.

4.1 Livello morfosintattico

L'analisi della composizione dei turni conversazionali ha evidenziato che i bambini del gruppo ASD producono un numero di unità inferiore rispetto al GC e si servono meno spesso di frasi vere e proprie, ossia di strutture sintattiche complete sul piano sintattico e interpretabili indipendentemente dal contesto.

Un aspetto rilevante, anche se non statisticamente significativo, è l'elevato numero di turni vuoti nel gruppo ASD. Poiché questi compaiono soprattutto in corrispondenza di risposte che l'interlocutore si aspetta di ricevere dal proprio *partner* comunicativo, ma che di fatto non vengono fornite, il dato offre importanti spunti di riflessione. Da una parte, suggerisce la presenza di anomalie nella «comunicazione ricettiva» (Vivanti, 2021: 108), ossia nella comprensione della comunicazione, particolarmente evidenti nei soggetti con sintomi più severi. Questi bambini faticano a riconoscere la fine del turno conversazionale del proprio interlocutore (cfr. Eigsti, De Marchena, Schuh & Kelley, 2011) e, di conseguenza, spesso, non sono in grado di rispondere alle domande, nemmeno quando esplicitamente sollecitati. Dall'altra parte, il mancato ricorso a strategie compensatorie (es. pause piene), che consentirebbero di guadagnare tempo e pianificare mentalmente la risposta, può essere interpretato come indice di un ridotto carico cognitivo e di uno scarso investimento comunicativo nell'interazione (cfr. Thurber, Tager-Flusberg, 1993). Questo riflette i deficit comunicativi e sociali che caratterizzano il quadro clinico dei soggetti con ASD (APA, 2013). In altre parole, la tendenza a non adottare *escamotage* comunicativi per evitare silenzi che potrebbero compromettere l'interazione segnala una limitata capacità di entrare in relazione con l'altro e una minore motivazione a partecipare attivamente alla conversazione.

Nelle produzioni del gruppo ASD, si è, inoltre, osservata una più alta tendenza a utilizzare strutture internamente poco articolate sotto il profilo sintattico. In particolare, è possibile isolare un sotto-gruppo di sei bambini con età diverse che si esprimono adottando perlopiù uno stile telegrafico e frasi nucleari, talvolta incomplete sia sul piano sintattico che su quello semantico. Stando alle valutazioni della GALS (Cipriani, Chilosi, Bottari & Pfanner, 1993; Chilosi, Pfanner, Villani & Bottari, 2002), ossia della Griglia di Analisi del Linguaggio Spontaneo, questo sotto-gruppo si situerebbe in uno stadio intermedio tra il livello presintattico (Livello 2) e il livello morfologico (Livello 4), che generalmente inquadrano le traiettorie di sviluppo seguite dai bambini normotipo di età compresa tra i 18 e i 32 mesi. Considerando che tutti i bambini osservati in questo studio hanno tra i 6 e i 13 anni, è evidente che, per il sotto-gruppo su menzionato, lo sviluppo morfosintattico non è conforme con i dati normativi e segue, dunque, un ritmo più lento di quanto atteso rispetto all'età. Questo dimostra, inoltre, che le competenze linguistiche e

comunicative esibite dai soggetti del gruppo ASD sono tra loro molto eterogenee e non sempre riconducibili a un «profilo» (Gagliardi, 2021) univoco.

L'analisi delle costruzioni frastiche ha evidenziato una ridotta complessità morfosintattica nelle produzioni del gruppo ASD rispetto al GC. Nella fattispecie, i bambini con ASD utilizzano prevalentemente frasi principali, mentre nel GC si osserva una distribuzione più equilibrata tra principali, coordinate e subordinate. Questo suggerisce che i bambini con autismo, tendenzialmente, si limitano a narrare gli eventi senza istituire relazioni logico-semantiche esplicite tra loro. Detto altrimenti, prevale nel gruppo ASD uno stile discorsivo formalmente frammentato e sconnesso. Per quanto riguarda le incidentali, pur essendo scarsamente presenti in entrambi i gruppi, risultano quasi assenti nelle produzioni del gruppo ASD. Questo dato è significativo poiché le frasi incidentali rappresentano una strategia che il soggetto adotta durante la conversazione per prendere tempo e proseguire lo scambio nella maniera più adeguata, oppure per mitigare l'atto comunicativo appena realizzato o ancora in procinto di realizzazione. La ridotta presenza di queste costruzioni nel gruppo ASD, con un solo esempio registrato, rafforza l'ipotesi già discussa di un limitato coinvolgimento nell'attività conversazionale, suggerendo una più generale incapacità di tenere in conto la prospettiva dell'interlocutore (cfr. Pfanner *et al*., 2008).

Anche l'analisi delle subordinate ha segnalato differenze significative tra i due gruppi (cfr. Tab. 4). In particolare, i controlli usano in media un numero più alto di avverbiali, argomentali e relative in confronto ai bambini con ASD. Inoltre, si è notato che, mentre i bambini del gruppo ASD usano quasi esclusivamente subordinate avverbiali con valore temporale, causale e – in modo più contenuto – finale, i controlli esibiscono una maggiore varietà, avvalendosi anche di consecutive, condizionali, avversative e comparative, del tutto assenti nelle produzioni del gruppo ASD. Per quanto riguarda le argomentali, invece, entrambi i gruppi usano principalmente le oggettive.

Rispetto alle subordinate relative, sebbene quelle sul soggetto siano padroneggiate da quasi tutti i bambini del campione, le relative sull'oggetto sono presenti nel GC ma inesistenti nel gruppo ASD. Questo dato suggerisce che, in confronto ai coetanei normotipo, i bambini con ASD seguano un ritmo più lento di acquisizione rispetto a determinate strutture morfosintattiche (Tager-Flusberg *et al*., 2005). Nella fattispecie, come già dimostrato da Durrleman, Hippolyte, Zufferey, Iglesias & Hadjikhani (2015) in merito alla comprensione delle frasi relative in adolescenti e adulti con ASD, la difficoltà insita nelle costruzioni relative sull'oggetto può essere dovuta alla coesistenza di due fenomeni che le rendono strutturalmente più complesse: da una parte, l'incassamento; dall'altra, la disposizione dei costituenti della frase in un ordine non-canonico (OSV) derivato dal procedimento del movimento sintattico, con cui l'oggetto viene spostato in posizione iniziale, sito più tipicamente occupato dal soggetto.

4.2 Livello pragmatico

Passando agli aspetti pragmatici, l'analisi della capacità di cogliere inferenze ha rivelato una differenza nelle *performance* tra i due gruppi, che, però, non trova riscontro sul piano statistico. Tuttavia, come evidenziato nella Tab. 5, mentre nel GC 16 bambini su 17 deducono correttamente il collegamento implicito tra i due eventi rappresentati nella sequenza di immagini somministrata, nel gruppo ASD il numero si riduce a 11. Va però precisato che, dei sei bambini con ASD che non colgono l'inferenza, tre si mostrano totalmente incapaci di anticipare un possibile scenario futuro (cfr. Adornetti, 2018: 133), mentre gli altri riescono, dopo diverse sollecitazioni, a pervenire alla soluzione. Nel GC, invece, l'unico bambino che non risulta in grado di cogliere il nesso tra i due eventi riesce, in un secondo momento, a fornire la risposta corretta.

Sul piano della coesione testuale, l'analisi relativa alla tipologia e all'appropriatezza delle espressioni referenziali (cfr. Tab. 6) ha mostrato che entrambi i gruppi si avvalgono principalmente di SND, Pron nulli e SNI, e che usano di frequente il SND in luogo del SNI per introdurre un personaggio o un elemento nuovo nel discorso, non ancora menzionato nella porzione di cotesto immediatamente precedente. Questo significa che entrambi i gruppi di bambini osservati non padroneggiano completamente le espressioni referenziali, almeno per quel che concerne l'oralità. Statisticamente, sono state rilevate delle differenze significative soltanto in merito a Pron atono, Pron relativo e SNsD. Quanto ai primi due, si è osservato che i controlli usano un numero di pronomi atoni e pronomi relativi superiore rispetto ai bambini con ASD. Un dato che correla con la maggiore complessità morfosintattica e il più consistente uso di frasi subordinate relative da parte del GC. Quanto al parametro SNsD, la sua significatività statistica è giustificata in parte dal fatto che, in più occasioni, i soggetti del gruppo ASD, soprattutto quelli che presentano un quadro comunicativo più compromesso, si esprimono tramite liste di nomi (cfr. es. 1); in parte dalla frapposizione di pause vuote tra i componenti di una sequenza o tra un turno e l'altro, che, spezzando il flusso del discorso, inducono il soggetto a ripristinarlo proponendo un sintagma spogliato del suo determinante (cfr. es. 2).

(1) S.F., 7.11
l'autobus / # **autobus** / # **autobus** / # <cade> nell'acqua //

(2) F.I., 7.2
il gatto / # mangia / **# torta** //

Un altro parametro indagato sul piano della connessità testuale è il numero di connettivi utilizzati nei testi dai bambini con ASD e dal GC. I risultati (cfr. Tab. 7) mostrano che i controlli producono in media un numero di connettivi superiore in confronto ai bambini con ASD. Considerando che i connettivi sono necessari ai fini della costruzione dell'ossatura di un testo che sia inerentemente coeso, questo dato conferma la mancanza di unitarietà, almeno dal punto di vista formale, nei discorsi elaborati dal gruppo ASD. Una mancanza a sua volta spiegata dal ricorso più limitato a coordinate e subordinate, i cui introduttori servono, di fatto, a

tenere unite le varie porzioni di un testo che sono tra loro relate sul piano logico e cronologico.

Per quanto attiene all'analisi sulla gestione del discorso riportato nel *Bus Story Test*, i risultati (cfr. Tab. 8) dimostrano che, sebbene nei due gruppi non si raggiunga mai punteggio pieno per nessuno dei quattro enunciati attesi, i bambini del GC riescono più spesso dei bambini con ASD a riportare in maniera fedele le parole dei personaggi. Quanto alla tipologia di discorso riportato, come illustrato nella Fig. 6, in entrambi i gruppi si rileva un uso consistente di discorso diretto, inserito quasi sempre tramite apposita cornice citante, eccetto quando il bambino si limita a rispondere alla domanda dell'intervistatore senza riprendere il *verbum dicendi*. Sebbene venga utilizzato meno del discorso diretto, sono state rintracciate in totale 20 occorrenze di discorso indiretto, espresso per mezzo di una proposizione subordinata, esplicita o implicita, dipendente da un verbo dichiarativo (es. "dire"). Di queste 20 occorrenze, soltanto 6 sono ascrivibili al gruppo ASD, mentre le restanti 14 sono realizzate dal GC. Ci sono però degli enunciati, prodotti da entrambi i gruppi, che meritano una particolare attenzione, dal momento che vengono inglobati direttamente nel flusso della narrazione senza che vi siano le tipiche spie morfosintattiche che ne preannunciano l'inserimento, ossia senza la presenza di clausola citante (Mortara Garavelli, 2001). Sulla base di queste caratteristiche, si è preferito parlare in questi casi di discorso indiretto libero, tramite cui i bambini ripropongono i pensieri e gli stati d'animo dei personaggi, fungendo da narratori onniscienti. Esempi di questo fenomeno sono illustrati di seguito.

(3) E.S., 13.5 (ASD)
poi [/1] poi l'autobus continuava a camminare sulla strada / **ma poi si scocciò di andare** / **# sempre nella strada** //

(4) R.D.S, 11.2 (GC)
poi se ne andò // # **ed era stufo di stare sempre in strada** / allora / # scavalcò un recinto / e mmh andò in campagna //

4.3 Livello della fluenza

Infine, anche sul piano della fluenza sono emerse differenze significative tra i due gruppi. Considerando i dati ottenuti dall'analisi condotta singolarmente su ogni prova, una prima osservazione riguarda la durata del *task*, il cui *p-value* raggiunge significatività statistica soltanto nella prova "*Cookie Theft*". Difatti, il gruppo ASD ha impiegato in media il doppio del tempo rispetto al GC per completare il *task* (cfr. Tab. 9). I bambini con ASD mostrano peraltro una forte eterogeneità interpersonale: per esaurire il compito, alcuni hanno impiegato pochi secondi, altri invece diversi minuti perché si perdevano continuamente nella descrizione di dettagli irrilevanti. Nel GC, al contrario, la situazione sembra nel complesso più bilanciata, in quanto i valori emersi dall'analisi risultano tra loro più coerenti.

Un dato molto interessante, che però non trova riscontro sul piano statistico in nessuna prova, riguarda la durata delle pause vuote, che risulta sempre più alta nel gruppo ASD rispetto al GC (cfr. Tab. 9, 10 e 11). Questo dato è importante nella

misura in cui contribuisce, nella prova "Una festa di compleanno" e nel "*Bus Story Test*", a decretare la disparità statisticamente rilevante individuata tra i due gruppi in merito al TPT. Questo parametro, infatti, rappresenta la durata complessiva dei turni conversazionali ascrivibili al soggetto osservato durante l'intero scambio comunicativo con l'intervistatore, a cui viene sottratta la durata complessiva delle pause vuote prodotte dal soggetto. È, quindi, evidente che la durata maggiore delle pause vuote nel gruppo ASD nelle due prove considerate determina un TPT, ossia un contributo verbale alla conversazione, mediamente inferiore a quello rilevato per il GC.

Particolarmente eloquenti sono i dati relativi al numero e alla durata delle pause piene, che di fatto raggiungono significatività statistica in tutte e tre le prove somministrate. Nella fattispecie, esse vengono utilizzate in modo più o meno consistente nel caso del GC e in maniera molto limitata nel gruppo ASD. Siccome le pause piene sono generalmente usate sia per prendere tempo e pianificare mentalmente quanto si vuole dire di seguito, sia per evitare turni vuoti che avrebbero un impatto negativo sull'esito dell'interazione in corso, il fatto che nel gruppo ASD esse siano ridotte permette, ancora una volta, di consolidare l'idea che nei soggetti con autismo vi sia un minore carico cognitivo e uno scarso investimento comunicativo nell'interazione (Thurber, Tager-Flusberg, 1993).

Un'ultima osservazione riguarda il numero di pause vuote rintracciate nel "*Bus Story Test*", che risulta maggiore nel GC rispetto al gruppo ASD, restituendo un *p-value* indicativo di significatività statistica (cfr. Tab. 11). Questo dato può essere spiegato in virtù della maggiore durata del *task* e del valore più elevato di TPT rilevati nel GC. Nel dettaglio, mentre i controlli hanno riportato la storia in tutti (o quasi tutti) i suoi passaggi, servendosi dunque delle pause vuote per ricordarli, buona parte dei bambini con ASD ha tralasciato una serie di informazioni più o meno necessarie. Di conseguenza, le pause vuote non sembrano aver avuto, in questo gruppo, il medesimo ruolo determinante individuato invece nel GC.

Per concludere, i dati emersi dalle tre analisi nel complesso (cfr. Tab. 12) mostrano che i parametri per cui si raggiunge significatività statistica sono essenzialmente il TPT e il numero e la durata delle pause piene.

5. *Conclusione*

Lo studio ha rintracciato differenze significative tra il gruppo ASD e il GC, sul piano sia quantitativo che qualitativo, rivelando un'interessante interazione tra i vari livelli di analisi indagati, che si esplica in *pattern* conversazionali e narrativo-discorsivi molto diversi per le due coorti.

In generale, le unità costitutive dei turni prodotti dal gruppo ASD risultano in media meno articolate sul piano morfosintattico, con un uso limitato di strutture frastiche complesse e una prevalenza di proposizioni principali rispetto a coordinate e subordinate. Questa tendenza ha influito negativamente sul piano della coesione testuale, poiché lo scarso ricorso a strategie di coordinazione e subordinazione ha

comportato, nel gruppo ASD, un uso più limitato di connettivi rispetto al GC. Di conseguenza, le produzioni dei bambini con ASD appaiono frammentate e poco coese, caratterizzate talvolta da un *focus* eccessivo sui dettagli a scapito delle informazioni necessarie per una comprensione efficace da parte dell'ascoltatore. Questo dato trova riscontro nel maggiore tempo impiegato dal gruppo ASD, rispetto ai controlli, per completare i tre *task* nel complesso.

Parimenti interessanti sono i risultati emersi in relazione alle frasi parentetiche. Trattandosi di strategie comunicative che consentono di portare avanti una conversazione e di soddisfare le necessità dell'interlocutore, la loro quasi assenza nelle produzioni dei bambini con ASD è stata interpretata come minore coinvolgimento nell'attività conversazionale e come una scarsa considerazione del *partner* comunicativo. Un'osservazione, questa, che ha trovato conferma nei dati ottenuti in merito al numero e alla durata delle pause piene, reputate anch'esse strategie fondamentali nel corso di un'interazione, perché permettono di prendere tempo e pianificare in modo adeguato quanto si intende dire di seguito. L'analisi sul piano della fluenza ha, infatti, dimostrato che i soggetti del gruppo ASD si servono mediamente di un numero di pause piene più contenuto rispetto ai controlli.

Le considerazioni sin qui fatte consentono di concludere che lo studio ha conseguito l'intento di rintracciare indici linguistici in grado di distinguere l'eloquio autistico da quello normotipo. Tuttavia, il numero esiguo di soggetti reclutati non consente di delineare un vero e proprio profilo linguistico-comunicativo del gruppo ASD, in quanto comportamenti osservati nel caso di pochi hanno avuto un ruolo determinante nel calcolo della media sul gruppo.

In ottica futura si ritiene pertanto fondamentale incrementare il campione, coinvolgendo eventualmente più centri riabilitativi attivi sul territorio campano. Si suggerisce, inoltre, di approfondire l'analisi morfosintattica, poiché questo studio dimostra che, oltre alla pragmatica, anche gli aspetti formali del linguaggio sono, in modo più o meno evidente, compromessi nei soggetti con ASD.

Riferimenti bibliografici

Adornetti, I. (2018). *Patologie del linguaggio e della comunicazione*. Roma: Carocci.

American Psychiatric Association (APA) (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth edition (DSM-5)*. Washington (D.C.) – London: American Psychiatric Publishing.

Arciuli, J., Brock, J. (Eds.) (2014). *Communication in autism*, 11. Amsterdam: John Benjamins Publishing Company.

Biancalani, S., Gagliardi, G. & Innocenti, M. (2023). Aspetti soprasegmentali e non verbali nel Disturbo dello Spettro Autistico: uno studio pilota. In Castegneto, M., Ravetto, M. (a cura di), *La comunicazione parlata. Spoken communication* (2020). Roma: Aracne, 829-854.

Cipriani, P., Chilosi, A.M., Bottari, P. & Pfanner, L. (1993). *L'acquisizione della morfosintassi in italiano – Fasi e processi*. Padova: Unipress.

Cipriani, P., Chilosi, A.M., Pfanner, L., Villani, S. & Bottari, P. (2002). Il ritardo del linguaggio in età precoce: profili evolutivi ed indici di rischio. In Caselli, M.C., Capirci, O. (a cura di), *Indici di rischio nel primo sviluppo del linguaggio. Ricerca, clinica, educazione*. Milano: Franco Angeli, 377-393.

Cresti, E., Moneglia, M. (2005). *C-ORAL-ROM. Integrated Reference Corpora for spoken Romance Languages*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Dennis, M., Lazenby, A.L. & Lockyer, L. (2001). Inferential language in high-function children with autism. In *Journal of autism and developmental disorders*, 31(1), 47-54.

Diehl, J.J., Bennetto, L. & Young, E.C. (2006). Story recall and narrative coherence of high-functioning children with autism spectrum disorders. In *Journal of abnormal child psychology*, 34(1), 83-98.

Durrleman, S., Hippolyte, L., Zufferey, S., Iglesias, K. & Hadjikhani, N. (2015). Complex syntax in autism spectrum disorders: A study of relative clauses. In *International Journal of Language & Communication Disorders*, 50(2), 260-267.

Eigsti, I.M., de Marchena, A.B., Schuh, J.M. & Kelley, E. (2011), Language acquisition in autism spectrum disorders: A developmental review. In *Research in autism spectrum disorders*, 5(2), pp. 681-691.

ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Gagliardi, G. (2021). *Che cos'è la linguistica clinica*. Roma: Carocci Editore.

Goodglass, Barresi, B., Weintraub, S., Segal, O., Goodglass, H. & Kaplan, E. (2001). *BDAE: the Boston Diagnostic Aphasia Examination* (3rd edition.). Lippincott Williams & Wilkins.

Malvessi Mittmann, M., Barbosa, P. (2016). An automatic speech segmentation tool based on multiple acoustic parameters. In *CHIMERA: Romance Corpora and Linguistic Studies*, 3(2), 133-147.

Moneglia, M., Cresti, E. (2006). C-ORAL-ROM. Prosodic boundaries for spontaneous speech analysis. In Kawaguchi, Y., Zaima, S., Takagaki, S. (Eds.), *Spoken Language Corpus and Linguistics Informatics*. Amsterdam: Benjamins, 89–112.

Mortara Garavelli, B. (2001). Il discorso riportato. In Renzi, L., Salvi, G. & Cardinaletti, A. (a cura di), *Grande grammatica italiana di consultazione*. Bologna: il Mulino, 429-470.

Naigles, L.R., Tek, S. (2017). "Form is easy, meaning is hard" revisited: (re) characterizing the strengths and weaknesses of language in children with autism spectrum disorder. In *Wiley interdisciplinary Reviews: Cognitive Science*, 8(4), e1438.

Padovani, R., Pattuzzi, L. (2017). Analisi della conversazione nell'autismo e nei disturbi della comunicazione sociale (pragmatica). Applicazione delle scale TOPICC e Pragmatic Rating Scale. In *Autismo e disturbi dello sviluppo*, 15(1), 97-115.

Paul, R., Cohen, D.J. (1984). Responses to contingent queries in adults with mental retardation and pervasive developmental disorders. In *Applied Psycholinguistics*, 5(4), 349-357.

Pfanner, L., Tancredi, R. & Marcheschi, M. (2008). Comunicazione e linguaggio nei disturbi pervasivi dello sviluppo. In *Giornale di Neuropsichiatria dell'Età Evolutiva*, 28, 59–74.

Renfrew, C. (2015). *Bus Story Test. A test of narrative speech*. Londra: Speechmark.

Sabatini, F., Camodeca, C. & De Sanctis, C. (2014). *Conosco la mia lingua*. Torino: Loescher.

Serianni, L., Della Valle, V. (2017). *Italiano. Le regole, le parole, i testi*. Milano: Edizioni Scolastiche Bruno Mondadori.

Tager-Flusberg, H., Paul, R. & Lord, C. (2005). Language and Communication in Autism. In Volkmar, F.R., Paul, R., Klin A. & Cohen D. (a cura di), *Handbook of autism and pervasive developmental disorders: Diagnosis, development, neurobiology, and behavior*. Hoboken, NJ: John Wiley & Sons Inc, 335–364.

Thurber, C., Tager-Flusberg, H. (1993). Pauses in the narratives produced by autistic, mentally retarded, and normal children as an index of cognitive demand. In *Journal of Autism and Developmental disorders*, 23(2), 309-322.

Vivanti, G. (2021). *La mente autistica. Le risposte della ricerca scientifica all'enigma dell'autismo* (2ª ed.). Firenze: Hogrefe Editore.

Wilkinson, K.M. (1998). Profiles of language and communication skills in autism. In *Mental retardation and developmental disabilities research reviews*, 4(2), 73-79.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J. & Tapias, D. (a cura di), *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. Paris: ELRA, 1556-1559.

Young, E.C., Diehl, J.J., Morris, D., Hyman, S.L. & Bennetto, L. (2005). The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders. In *Language, Speech, and Hearing Services in Schools*, 36, 62-72.

DOI: 10.17469/O2111AISV000014

VIRGINIA CHITI, VALENTINA SACCONE, ALESSANDRO PANUNZI

L'alterazione degli aspetti pragmatici e prosodici nel Disturbo dello Spettro Autistico: analisi del parlato di bambini in età scolare[1]

Pragmatic and prosodic features in Autism Spectrum Disorder: analysis of the speech of school-aged children

The paper aims to investigate pragmatics and prosody in Autism Spectrum Disorder, always impaired following the literature. A pilot study was conducted on a sample of semi-spontaneous speech of 9 children with ASD (8-12y), searching which parameters reflect their perceptive atypia. After an overview of the state of the art, we present a multilevel analysis of acoustic parameters, lexical features, and information structure. The variation of the F0 range together with intensity describes the prosodic trend of the sample along the axis of variation between very monotonous and strongly emphatic voice; atypical types of word-fragmentation emerge; a wider fragmentation leads to a telegraphic speech style alternatively to a drop-down list of utterances. Hence, inside the heterogeneity of the sample, a cluster of features emerge framing what is perceived as atypical in the prosody of ASD speech and the pragmatic level of discourse structure.

Keywords: Autism Spectrum Disorder, prosody, pragmatics, spontaneous speech.

1. *Introduzione*

Il presente contributo si propone di indagare la pragmatica e la prosodia del linguaggio di bambini nello Spettro Autistico a partire dalla considerazione, evidente nella realtà clinica e confermata dalla letteratura, che l'alterazione di queste due dimensioni comunicativo-linguistiche si configura come una costante, nonostante l'eterogeneità interindividuale intrinseca nella definizione stessa del disturbo. Come primo passo, è stato condotto uno studio pilota finalizzato ad indagare le alterazioni pragmatico-prosodiche rilevabili dall'analisi del parlato di bambini con ASD per inquadrare il fenomeno in maniera strutturata e far luce su un panorama teorico ad oggi molto sfumato. Scopo generale dello studio pilota è infatti gettare le basi per definire un quadro di variazione del fenomeno in vista della raccolta di un corpus più ampio che possa poi fornire un effettivo contributo alla pratica clinica logopedica. In tal senso, gli obiettivi specifici che qui affrontiamo sono: fornire dei materiali e una metodologia di analisi della prosodia nell'ASD; ricercare specifici parametri

[1] Gli autori hanno concepito insieme il lavoro. Nello specifico, Virginia Chiti ha scritto i paragrafi 1-3, Valentina Saccone i paragrafi 4-5, e Alessandro Panunzi ha coordinato la ricerca.

che riflettano la percettiva atipia del parlato; sensibilizzare la comunità clinica nei confronti della componente prosodica del linguaggio, ad oggi considerata in modo marginale rispetto alle altre aree linguistiche nella ricerca e nella pratica logopedica.

Per un inquadramento dello stato dell'arte, presentiamo di seguito le principali caratteristiche dell'ASD, concentrandoci sui deficit nella sfera comunicativo-linguistica e le atipie riscontrate in letteratura in ambito pragmatico-prosodico, per poi descrivere nel dettaglio il caso di studio e le analisi condotte sul campione.

2. *Il Disturbo dello Spettro dell'Autismo*

Il Disturbo dello Spettro dell'Autismo è identificato nel DSM-5 (American Psychiatric Association, 2013) come una sottocategoria dei disturbi del neurosviluppo: trattasi quindi di una sindrome comportamentale, con esordio nel periodo precoce dello sviluppo, spiegata, seppur in maniera ancora oggi poco definita, da un'alterazione dello sviluppo del sistema nervoso. Le principali aree sintomatologiche riguardano la sfera comunicativa e di interazione sociale e la ristrettezza comportamentale e di interessi che determinano, in generale, una compromissione più o meno importante del funzionamento del soggetto.

Le alterazioni relative all'area della comunicazione e dell'interazione sociale si configurano come una costante e si esprimono in vario modo in tutta "una serie di comportamenti, tra cui un approccio sociale anomalo, fallimento nella reciprocità della conversazione, ridotta condivisione di interessi, emozioni, sentimenti, incapacità di iniziare o rispondere a interazioni sociali; deficit dei comportamenti comunicativi non verbali utilizzati nelle interazioni sociali, anomalie del contatto visivo, e del linguaggio del corpo o deficit della comprensione e uso dei gesti, o mancanza totale di espressività facciale e di comunicazione non verbale" (ISAAC Italy, Società Internazionale Comunicazione Aumentativa e Alternativa, 2017). Per definizione quindi si riscontra una generica alterazione delle capacità comunicative e, non a caso, principalmente e universalmente compromessa nell'ASD risulta quella sfera linguistica relativa all'abilità di adattare il linguaggio al contesto di comunicazione, ovvero la pragmatica (Tager-Flusberg, Paul, Lord, Volkmar, Paul, Klin, 2005; Young, Diehl, Morris, Hyman, Bennetto, 2005).

La seconda area di funzionamento alterato riguarda la sfera comportamentale: si osserva la presenza caratteristica in questi soggetti di una certa rigidità che si esprime tramite ristrettezza e reiterazione, ripetitività di comportamenti, attività ed interessi e nella pratica viene manifestata con "movimenti, uso degli oggetti o eloquio stereotipati o ripetitivi; aderenza alla routine priva di flessibilità o rituali di comportamento verbale o non verbale; interessi molto limitati, anomali per intensità o profondità; iper- o iporeattività in risposta a stimoli sensoriali o interessi insoliti verso aspetti sensoriali dell'ambiente" (DSM-5, American Psychiatric Association, 2013).

Volendo approfondire dal punto di vista teorico la sfera comunicativo-linguistica, si ritiene rilevante specificare uno dei concetti principali e distintivi

dell'Autismo: l'eterogeneità interindividuale. L'estrema variabilità clinica associata all'ASD spiega l'introduzione del termine "spettro" nell'ultima edizione del DSM ed è riferita in generale a tutti i livelli di funzionamento di un individuo, pertanto anche, e soprattutto, alle competenze comunicativo-linguistiche. Se però le difficoltà di comunicazione rappresentano una peculiarità imprescindibile del disturbo con una frequente dissociazione tra forma, contenuto ed uso del linguaggio (Barone, Bravaccio, Muratori, Salomone, Venuti, Vicari, 2020), proprio in ragione del concetto di spettro, queste possono declinarsi in vario modo ed interessare diversi domini linguistici. Sotto la stessa etichetta diagnostica rientrano soggetti che parlano fluentemente, soggetti con linguaggio telegrafico o ridotto a poche parole e soggetti, addirittura, completamente non verbali.

La conseguenza di questa eterogeneità clinica non rende pertanto possibile delineare un profilo comunicativo-linguistico caratteristico dei soggetti con ASD. Nonostante questo, sono state identificate alcune atipicità comunicativo-linguistiche più frequentemente associate a questo disturbo (la cui presenza isolata però non è di per sé patognomonica).

2.1 Sfera comunicativo-linguistica nell'ASD

Già fin dai primi momenti di vita, nei bambini con Autismo, si osserva più facilmente e in modo più o meno evidente, la compromissione della comunicazione e di quella naturale propensione alla relazione e alla socialità che si evidenza nello sviluppo normotipico. Si riscontrano varie manifestazioni di scarso interesse verso la socialità e verso le persone che li circondano, come per esempio la presenza di uno sguardo sfuggente che tende ad orientarsi verso gli oggetti, la presenza di modalità comunicative non finalizzate alla condivisione quanto più al raggiungimento di un obiettivo, la presenza di una gestualità caratterizzata principalmente da azioni ritualizzate e gesti strumentali, che indicano la natura prevalentemente "richiestiva" degli atti comunicativi e di interazione che questi bambini esibiscono (caratteristico è l'uso protesico dell'adulto), la mancanza o la riduzione di gesti referenziali (D'Amico, Devescovi, 2013).

Dal punto di vista dello sviluppo strettamente linguistico, invece, è a partire dal secondo anno di vita che si definisce ancor di più l'eterogeneità che caratterizza il linguaggio dei bambini con Autismo. In linea generale, si assiste ad un ritardo di linguaggio importante che poi esiterà in vario modo: ci sarà chi non lo svilupperà mai, chi lo svilupperà molto in ritardo e sarà estremamente limitato sia dal punto di vista della forma che del contenuto, chi lo svilupperà più in termini di contenuto e meno di forma, chi svilupperà un linguaggio formalmente corretto ma del tutto non funzionale. Caratteristica è, inoltre, la regressione delle capacità: nel secondo anno di vita, circa il 15-20 % dei bambini che rientrano nello spettro autistico subisce una regressione funzionale che coinvolge le abilità linguistico-comunicative. In linea generale, le capacità di interazione e di responsività continuano ad essere ampiamente atipiche e deficitarie (Barone *et al.*, 2020).

Secondo lo sviluppo normotipico, dai 3 anni fino all'ingresso alla scuola primaria, il linguaggio continua la sua espansione diventando ancora più complesso e articolato in termini di forma e contenuto: cresce la complessità grammaticale, compaiono nuove tipologie di frasi, la lunghezza media dell'enunciato aumenta, l'intellegibilità migliora, si sviluppano le abilità narrative con l'acquisizione di migliori capacità di orientamento spazio-temporale e di astrazione. Le ricerche sull'Autismo mostrano come, in età prescolare, i cambiamenti evolutivi delle capacità comunicative siano collegati al generale livello di sviluppo e alla gravità della sintomatologia. Come nello sviluppo tipico, anche nell'Autismo la comunicazione non verbale risulta predittiva del successivo sviluppo linguistico: in particolare, attenzione condivisa e intenzionalità comunicativa di tipo dichiarativo si configurano come i migliori predittori dello sviluppo del linguaggio; quindi, anche tra i bambini con Autismo coloro che avranno mostrato migliori abilità di comunicazione preverbale, seppur con i limiti che il disturbo stesso comporta, avranno probabilmente maggiori potenzialità in termini di sviluppo linguistico (D'Amico, Devescovi, 2013; Barone *et al.*, 2020).

Nel linguaggio precoce dei bambini con ASD che sviluppano un linguaggio verbale, al di là delle differenze idiosincratiche, la letteratura evidenzia come caratteristiche ricorrenti l'ecolalia, ossia la ripetizione immediata o ritardata di parole e frasi pronunciate da qualcun altro, le difficoltà nella coniugazione verbale (spesso c'è una resistenza all'utilizzo dei verbi all'infinito) e nell'uso corretto dei pronomi personali e delle particelle pronominali (Barone *et al.*, 2020).

Infine, sempre in riferimento a quella porzione di soggetti con ASD per i quali ha senso parlare di un'evoluzione, seppur differente, delle capacità comunicativo-linguistiche, soprattutto a partire dall'inizio della scolarizzazione emerge in modo più evidente la compromissione della sfera pragmatica del linguaggio. Se nel corso dell'età scolare e dell'adolescenza i soggetti a sviluppo tipico gradualmente sviluppano un linguaggio più complesso e ricco di termini astratti o più tecnici, per lo più appresi a scuola, e sviluppano un linguaggio più flessibile in termini di adeguamento al contesto comunicativo, i soggetti con ASD possono presentare notevoli difficoltà linguistiche che si ripercuotono notevolmente sull'interazione e la socializzazione con i coetanei, soprattutto nell'adolescenza (Barone *et al.*, 2020). Dai sei anni in poi è frequente riscontrare la produzione di frasi bizzarre, la perseverazione linguistica nelle domande, la mancanza di intenzionalità comunicativa, i disturbi di fluenza e permane come costante l'alterazione della prosodia (Brambilla, 2011). Inoltre, tra le maggiori e più frequenti difficoltà, si menzionano quelle nella comprensione e nell'uso di espressioni linguistiche e forme di comunicazione che prevedono un certo livello di astrazione, come nel capire ed usare il sarcasmo, i modi di dire, le metafore, le battute; si riscontrano inoltre difficoltà nei processi di deduzione e di mentalizzazione. Tutto questo molto spesso conduce ad una comprensione e produzione linguistica caratteristicamente letterali e cioè strettamente circoscritte e limitate al significato diretto delle parole usate, da cui deriva talvolta una bizzarra inappropriatezza contenutistica. Oltre a queste,

si riscontrano difficoltà legate al rispetto delle regole sociali di conversazione, per esempio un mancato rispetto del turno, incapacità d'ascolto, non attinenza al tema di conversazione, incapacità di inibire osservazioni irrilevanti, estranee o inappropriate all'argomento di comunicazione, o anche con l'assenza delle forme di cortesia (Barone *et al.*, 2020). Si capisce bene come tutte queste limitazioni e difficoltà conducono spesso e volentieri allo sviluppo di un linguaggio che, seppur formalmente corretto, risulta molto rigido e pragmaticamente inadeguato.

2.2 La Prosodia nell'ASD

Se è vero quindi che nell'ASD il linguaggio può essere estremamente variegato e presentare deficit di vario genere che si manifestano in una casistica molto eterogenea, da casi di completa assenza di comunicazione verbale fino addirittura a stili linguistici morfologicamente adeguati e con un lessico ricco e "ampolloso", è altrettanto vero e confermato dagli studi e soprattutto dalla realtà clinica e quotidiana, nonché fulcro e punto di partenza del presente studio, che all'interno di questa enorme variabilità interindividuale, una sfera linguistica si presenta costantemente alterata in tutti gli individui con questo disturbo: la prosodia (Eigisti, Schuh, Mencl, Schultz, Paul, 2011).

Purtroppo, la letteratura restituisce poche informazioni e punti fermi nella descrizione della prosodia di soggetti con ASD. Gli studi sul tema, oltre ad essere scarsi e difficilmente confrontabili, di fatto sottolineano in modo piuttosto univoco l'impossibilità di definire un profilo prosodico patologico caratteristico del disturbo.

Ciononostante, è premessa comune che la prosodia sia una componente comunicativo-linguistica tipicamente e costantemente alterata nei soggetti con Autismo. Già Kanner (1943), nelle sue prime descrizioni, includeva tra le caratteristiche peculiari del disturbo un'insolita prosodia espressiva, insieme a difficoltà pragmatiche (Eigisti *et al.*, 2011; McAlpine, Plexico, Plumb, Cleary, 2014): nel relazionarsi con soggetti con ASD si percepisce di fatto sempre una certa "stranezza" ed evidenti anomalie e atipie delle competenze comunicativo-relazionali. Gli studi successivi descrivono soltanto in termini generici la prosodia dei soggetti con Autismo, come "monotona, robotica e pedante, non modulata nell'intonazione (piatta o al contrario esagerata) e nella rapidità di eloquio (troppo lenta o, viceversa, eccessivamente accelerata)" (Testi Padovani, Bilancia, Barbi, Musetti, 2015; Eigisti *et al.*, 2011), sottolineando che i deficit prosodici generalmente permangono anche quando altre aree linguistiche migliorano (Eigisti *et al.*, 2011; McCann, Peppé, 2003).

3. *Lo studio pilota*

Al fine di approfondire le caratteristiche prosodiche e pragmatiche dell'ASD nei bambini, è stato raccolto un campione di audio-registrazioni di parlato di bambini con ASD, per un totale di 150 minuti di registrazione. La ricerca è avvenuta in

collaborazione con la Fondazione MAiC di Pistoia, presso cui sono stati raccolti i dati nei mesi di luglio e agosto 2022[2].

Grazie all'aiuto degli operatori della Fondazione, è stato individuato un gruppo di soggetti il più omogeneo possibile in termini di sviluppo linguistico, età e funzionamento, secondo i seguenti criteri di selezione: diagnosi di Disturbo dello Spettro Autistico; presenza di linguaggio verbale intellegibile; lingua italiana come L1; frequentanti la scuola primaria o la scuola secondaria di primo grado[3].

È stato così selezionato un campione di 9 bambini[4], 7 maschi e 2 femmine[5], di età compresa fra gli 8 e i 12 anni.

Con l'obiettivo di garantire la riservatezza delle informazioni personali, ciascun bambino e la relativa registrazione sono stati identificati tramite un codice alfanumerico costituito da due elementi: un numero assegnato sequenzialmente sulla base dell'ordine di registrazione da 1 a 9, e la lettera M o F per indicare il sesso del soggetto. Si veda in Tab. 1 un riepilogo del campione raccolto:

Tabella 1 - *Riepilogo partecipanti*

ID	*Età*[6]	*Classe frequentata*[7]
1M	10A 11M	SP, classe V
2M	9A 1M	SP, classe III
3F	8A 4M	SP, classe II
4F	10A 11M	SP, classe V
5M	11A 10M	SM, classe I
6M	11A 10M	SM, classe I
7M	12A 4M	SM, classe I
8M	11A 8M	SM, classe I
9M	12A 3M	SM, classe I

[2] La ricerca è stata svolta nel contesto del progetto di tesi in Logopedia di Virginia Chiti presso l'Università degli Studi di Firenze (Chiti, 2022).

[3] Il range di età del campione risulta abbastanza ampio; si precisa però che tale scelta è stata necessariamente condizionata dalla disponibilità dell'utenza al momento della raccolta, nonché subordinata alla priorità di creare un gruppo di ricerca costituito da soggetti con livelli di funzionamento affini.

[4] Il campione era inizialmente composto da dieci soggetti; uno è poi stato escluso a causa della rilevante influenza della lingua albanese (lingua madre della famiglia della bambina) e di un livello di intellegibilità notevolmente inferiore rispetto al resto del campione.

[5] La distribuzione di genere riscontrata nel campione ricalca l'andamento della popolazione con ASD: in Italia i maschi risultano 4,4 volte più a rischio rispetto alle femmine (Ministero della Salute 2023).

[6] Sono riportate le età dei partecipanti al momento della registrazione, relativamente quindi al periodo compreso tra il 26 Luglio 2022 e il 5 Agosto 2022, nel formato A=anni; M=mesi.

[7] La classe frequentata è da riferirsi alla classe appena conclusa (SP=scuola primaria; SM=scuola secondaria di primo grado).

Le audio-registrazioni, della durata complessiva di circa 15 minuti ciascuna, sono state svolte durante le sedute di terapia settimanale, con il consenso informato dei genitori dei partecipanti; il contesto comunicativo-ambientale, di tipo semi-strutturato, è rimasto invariato per tutti e 9 i bambini, per ricreare le medesime condizioni emotive e comunicative. Gli incontri sono stati così strutturati: dopo aver presentato al bambino cosa sarebbe stato fatto durante la seduta, sono stati registrati per ogni partecipante una presentazione personale, la lettura di una storia per immagini e un momento finale di gioco. Si tratta quindi di parlato semi-spontaneo, registrato con microfono visibile[8].

Nella parte di presentazione personale è stato chiesto al partecipante di presentarsi e raccontare spontaneamente qualcosa di sé; in caso di verbalizzazione ridotta, i partecipanti sono stati incoraggiati con dei suggerimenti e/o domande generici relative alla classe frequentata, allo sport praticato e alle vacanze estive.

Per la "lettura" di una storia figurata, è stato utilizzato il test elaborato a partire dalla storia *Frog, Where are you?* (Mayer, 1969), che costituisce nella pratica clinica logopedica un test per la valutazione delle competenze narrative e rientra tra i cosiddetti test di "telling": fornendo un input visivo (nel caso specifico costituito da 24 immagini in sequenza che mostrano la breve avventura di un bambino che, col suo cagnolino, va alla ricerca della sua rana, fuggita durante la notte) è stato chiesto ai bambini osservare le immagini e contemporaneamente elaborare una narrazione, senza un precedente racconto da parte dell'esaminatore. Anche in questo contesto, alcuni bambini hanno richiesto alcune sollecitazioni, aiuti o correzioni e talvolta sono stati sollecitati a notare alcuni particolari.

Il momento di gioco è stato inserito per rendere più spontanea l'interazione e per creare uno stato emotivo positivo; ogni partecipante è stato invitato a scegliere liberamente un gioco di sua preferenza, spiegarne le regole e poi attivamente giocarci. Sono stati utilizzati giochi da tavolo di vario tipo, causa-effetto per i bambini più piccoli, di ragionamento per i più grandi.

La fase di raccolta dati si è conclusa con un momento di discussione con l'operatore che aveva in carico il bambino, per raccogliere informazioni anagrafiche e indicazioni sulla storia linguistica del partecipante, con riferimento particolare all'esposizione ad altre lingue e/o dialetti.

4. *Analisi*

4.1 Metodologia

Le registrazioni sono state trascritte e anonimizzate, secondo i criteri CHAT-LABLITA (MacWhinney, 2000; Moneglia, Cresti, 1997; Cresti, Moneglia, 2005; Cresti, Panunzi, 2013); il formato ortografico include indicazioni di tipo prosodico, segnalazioni di interruzioni e sovrapposizioni fra parlanti, disfluenze ed esitazioni

[8] Per le registrazioni è stato utilizzato un registratore portatile Zoom H6 con capsula microfonica XYH-6, composta da due microfoni unidirezionali inclinati di 90°.

(false partenze, frammentazioni di parola); contestualmente sono stati allineati testo e suono tramite l'uso del software di analisi vocale Praat (Boersma, Weenink, 2022) che ha permesso di svolgere in modo semi-automatico l'analisi prosodica del campione raccolto.

Come passo preliminare, è stata effettuata la segmentazione degli audio in enunciati e unità prosodico-informative, tramite l'identificazione dei confini prosodici terminali e non terminali (Teoria della Lingua in Atto, Cresti, 2000; Cresti, Moneglia, 2010; Izre'el, Mello, Panunzi, Raso, 2020) all'interno del continuum sonoro (indicati negli esempi rispettivamente con // e /). L'approccio scelto identifica come unità di riferimento pragmatica del parlato l'enunciato, portatore di forza illocutiva. Il riconoscimento dei confini ha permesso di individuare gli enunciati formati da un'unica unità prosodico-informativa, enunciati semplici, distinguendoli da quelli che invece sono formati da più unità prosodico-informative, ovvero gli enunciati complessi, che presentano al loro interno confini prosodici non terminali. Dagli audio sono stati scelti gli estratti analizzabili dal punto di vista acustico, in assenza di rumori, risate o sovrapposizioni fra parlanti, e per ogni enunciato così selezionato sono state ricavate misurazioni secondo più livelli di analisi quantitativa:

– misurazioni acustiche riguardo durata e velocità d'eloquio (sillabe/s), frequenza fondamentale (F0 minima, massima, media, range) e sforzo vocalico (intensità media; variazione di intensità);
– misurazioni lessicali come numero di tokens per registrazione, lunghezza media degli enunciati in termini di parole (LME), la presenza di frammentazioni di parola causate sia da esitazioni che da fenomeni particolari;
– misurazioni di struttura informativa degli enunciati quali la frequenza di enunciati semplici e complessi, la frequenza di segnali discorsivi e riempitivi.

Per una descrizione qualitativa delle registrazioni, sono stati annotati i segmenti di registrazione in cui spiccavano a livello percettivo fenomeni di qualità vocale atipici, come nei casi di *creaky voice*[9].

Dati quantitativi e note qualitative sono stati raggruppati per ogni bambino in una scheda informativa, che ne descrive l'eloquio e permette di delineare un profilo personale: sono qui ricapitolate abilità comunicative (responsività, stile di parlato), qualità vocale ed eventuali fenomeni espressivi particolari ritenuti rilevanti. I dati raccolti permettono di effettuare un confronto fra i vari partecipanti per mettere in luce eventuali similarità e differenze nel loro comportamento prosodico-pragmatico.

Le analisi sono state effettuate sulle porzioni di presentazione personale e *Frog Story* dei 9 bambini; il momento di gioco è stato lasciato da parte perché è risultato troppo rumoroso e quindi poco adatto alle misurazioni acustico-prosodiche. In totale quindi sono stati oggetto di analisi circa 70' di registrazioni (4910 tokens), come riassunto in Tab. 2.

[9] Non sono ancora state implementate le analisi acustiche sulla qualità della voce.

Tabella 2 - *Riepilogo dataset analizzato*

	minuti analizzati	*tokens (del partecipante)*
1M	07:05	390
2M	08:10	481
3F	08:00	445
4F	07:40	743
5M	05:35	466
6M	08:00	733
7M	08:25	619
8M	07:10	543
9M	07:35	490
Tot.	01:07:45	4910

Le registrazioni raccolte sono caratterizzate, già a un primo ascolto, da estrema varietà: si riscontra in alcuni bambini forte monotonia e in altri grande teatralità, voci gravi ma anche al contrario molto acute. Obiettivo della ricerca è stato quindi ricercare quali parametri permettano di descrivere meglio tali caratteristiche con la loro variazione e co-variazione. Presentiamo di seguito i risultati ottenuti per i tre livelli di analisi sopra individuati, integrando dove necessario la metodologia specifica.

4.2 Risultati

4.2.1 Misurazioni acustiche

Tramite il software Praat, i *tool* del programma e *script* personalizzati, sono state ottenute misurazioni relative al pitch e all'intensità del segnale acustico enunciato per enunciato. Per l'altezza tonale sono stati calcolati minimo, massimo, range e media di F0 in Hz e in st; per l'intensità, la media e un coefficiente (cvint) che permette di osservare la variazione di intensità rispetto al valore medio di ciascun parlante[10]; come indice di velocità d'eloquio, si è calcolato il rapporto fra numero totale di sillabe e tempo di enunciazione[11]. I dati sono stati poi raccolti parlante per parlante (Tabella 3[12]).

[10] Il coefficiente di variazione dell'intensità (cvint) è stato calcolato per ogni enunciato con la formula 100*sd[intensità]/media[intensità] (Barbosa, 2019); tale calcolo prescinde dalla distanza dal microfono e permette una misurazione più attendibile rispetto al dato dell'intensità media. Per questo motivo, si è scelto di considerare solo il cvint e non l'intensità media per confrontare fra loro i 9 bambini. Analisi in corso mirano ad integrare i valori di enfasi spettrale per rendere conto dello sforzo vocalico nella produzione parlata.

[11] Si intende per tempo di enunciazione la durata totale degli enunciati, comprensiva delle pause interne (fra unità prosodiche-informative) ma non delle pause fra enunciati (e quindi anche d'attacco o fine turno del parlante).

[12] Per il calcolo dei valori medi di F0 e intensità medie, le misurazioni enunciato per enunciato sono state pesate sulla durata di ogni segmento.

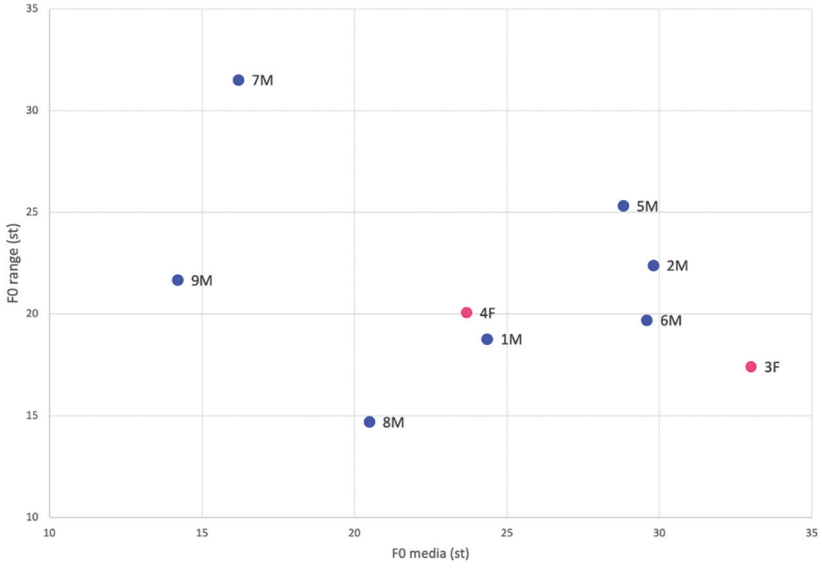
Tabella 3 - *Riepilogo delle misurazioni acustiche*

	F0 Range (Hz)	*F0 Range (st)*	*F0 media (Hz)*	*F0 media (st)*	*I media (dB)*	*cvint*	*velocità (sill/s)*
1M	263,057	18,770	204,100	24,351	64,504	9,148	4,383
2M	453,552	22,390	279,577	29,799	71,783	9,313	4,414
3F	350,223	17,410	336,719	33,019	64,673	12,721	3,363
4F	263,130	20,060	196,258	23,673	72,625	9,639	5,332
5M	418,477	25,320	264,169	28,818	72,444	8,978	5,061
6M	304,159	19,700	276,173	29,587	70,989	8,846	5,834
7M	300,090	31,510	127,393	16,191	71,639	8,350	5,349
8M	160,835	14,700	163,298	20,490	69,517	9,908	6,255
9M	159,958	21,670	113,586	14,205	71,710	7,590	7,566

Trattiamo in primo luogo le misurazioni relative al pitch, per poi osservare anche intensità e velocità d'eloquio.

Come si può osservare nel piano in Fig. 1[13], le misurazioni di F0 media (lungo l'asse delle ascisse) delineano come punti estremi del campione 9M e 3 F; l'età e il sesso dei bambini influenzano la distribuzione, in quanto 9M e 7M sono i più grandi del campione mentre 3F è la più piccola. Quando ad alti valori di F0 media si associano anche range maggiori (lungo l'asse delle ordinate), le voci sono percepite come più enfatiche (5M) o cantilenanti e con variazioni improvvise (2M).

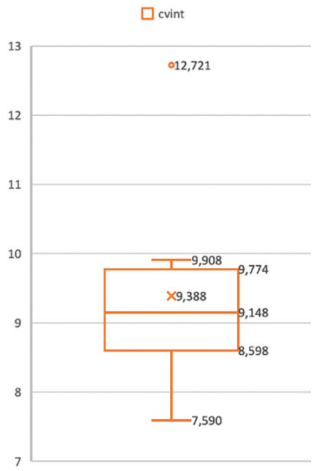
Figura 1 - *Misurazioni di F0 di semitoni, media e range*



[13] Nei grafici che seguono, sono stati distinti i colori in modo iconico rispettivamente blu/rosa per M/F, per avere un rapido colpo d'occhio sul comportamento dei partecipanti.

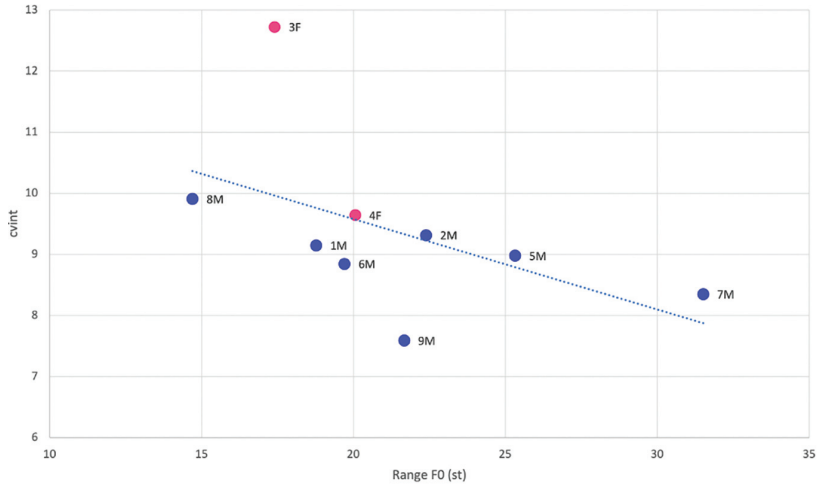
Il coefficiente cvint permette di illustrare un ulteriore fattore della percepita monotonia ovvero la scarsa variazione in intensità: a bassi valori di cvint corrisponde un'intensità più costante. Come mostrato in Fig. 2, sono anche in questo caso isolabili agli estremi 9M (cvint: 7,59) e 3F (cvint: 12,72), in cui il secondo si distacca nettamente dalla distribuzione del campione. Escludendo i due estremi, rispetto a questo parametro il campione risulta alquanto omogeneo, allontanandosi poco dal valore medio (indicato in figura dalla "x").

Figura 2 - *Distribuzione dei valori di cvint*



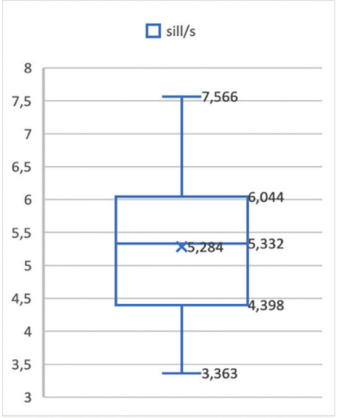
Combinando il parametro di range di F0 con il coefficiente di variazione dell'intensità (cvint) (Fig. 3), emergere una regressione lineare del campione che evidenzia come andamenti particolari quello di 3F e di 9M, i punti più lontani dalla distribuzione indicata dalla linea tratteggiata blu in figura. A questi corrisponde una percezione di maggiore atipia della voce: dalla più marcatamente monotona (9M) a quella più variata (3F), rispettivamente il punto più estremo verso il basso e verso l'alto, ai poli opposti di quello che potremmo individuare come un asse della monotonia vocale. I soggetti restanti sembrano compensare una mancata variazione in pitch con l'aumento della variazione di intensità. Al di sotto della linea, si posizionano i soggetti dal parlato più piatto; al di sopra quelli più enfatici. Maggiore è la distanza dalla linea, maggiore risulta anche la marcatezza percepita nella prosodia. L'andamento così individuato è indipendente dai valori di F0 che più possono essere soggetti a variazione interindividuale al di là delle atipie prosodiche.

Figura 3 - *Range F0 e cvint*



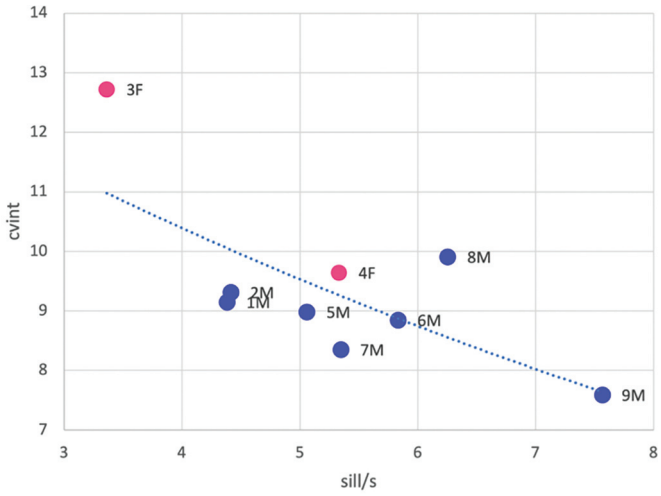
Osserviamo ora in Fig. 4 la distribuzione del campione secondo il parametro di velocità d'eloquio, misurata in sillabe/s. Anche in questo caso, 3F (3,36 sill/s) e 9M (7,57) si posizionano agli estremi opposti, inversamente però a quanto visto nelle precedenti distribuzioni. Come già osservato per il cvint, escludendo i due estremi il campione risulta alquanto omogeneo e i valori si concentrano intorno alla media (indicata in figura dalla "x"). Da analisi qualitative a campione risulta che in alcuni soggetti la velocità d'eloquio sia poco costante all'interno degli enunciati, creando la percezione di una velocità atipica che non può essere colta da un calcolo medio per parlante. Future analisi potranno approfondire l'argomento, osservando più nel dettaglio l'alternarsi da una parte di segmenti di parlato ipoarticolato e accelerato e dall'altra di sillabe con forti allungamenti vocalici in posizioni non attese, non dovute pertanto a ragioni pragmatico-informative.

Figura 4 - *Distribuzione dei valori di velocità d'eloquio, misurata in sillabe/s*



Confrontando i parametri di cvint e velocità d'eloquio (Fig. 5), sembra emergere una lieve tendenza per cui al crescere del primo parametro diminuisce il secondo, per cui i bambini che parlano più velocemente tendono ad esprimersi con una prosodia più piatta. 8M è il soggetto che più si discosta dall'andamento del gruppo.

Figura 5 - *Velocità d'eloquio (sill/s) e cvint*



4.2.2 Misurazioni lessicali

Per quanto riguarda l'analisi lessicale, abbiamo misurato la dimensione enunciato secondo il numero di parole e analizzato particolari fenomeni di frammentazione che vanno al di là degli episodi di parole ritrattate, auto-/etero-interruzioni o disfluenze del parlato spontaneo. I dati sono raccolti nella tabella sottostante (Tab. 4).

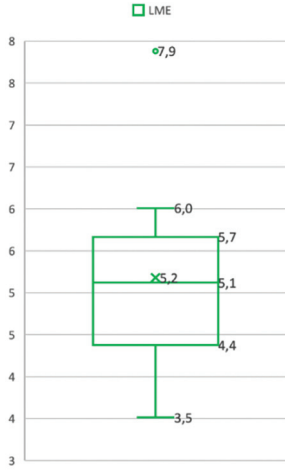
Tabella 4 - *Riepilogo dei dati lessicali*

	Tokens	*Enunciati*	*LME*
1M	390	111	3,5
2M	481	99	4,9
3F	445	99	4,5
4F	743	144	5,2
5M	466	91	5,1
6M	733	93	7,9
7M	619	103	6,0
8M	543	102	5,3
9M	490	115	4,3

Il parametro di Lunghezza Media degli Enunciati (LME) è calcolato dividendo il numero di tokens non ritrattati per il numero di enunciati; riflette la complessità

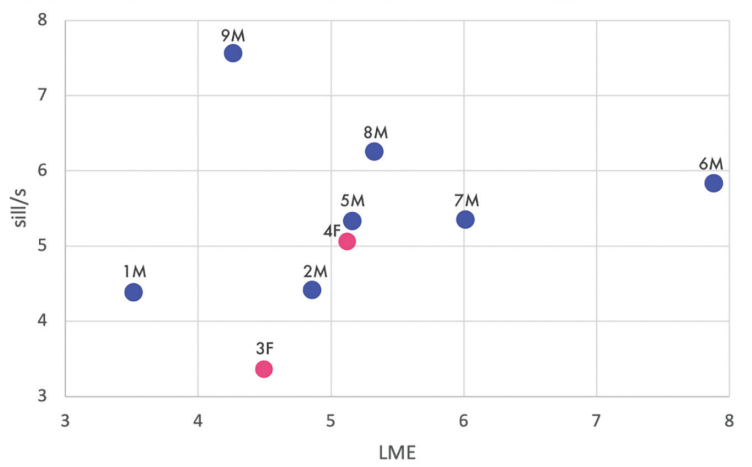
delle strutture del parlato rispetto al numero di parole per sequenza prosodicamente terminata e pragmaticamente autonoma. Il dato varia ampiamente nel campione in esame, da un massimo di 7,9 a un minimo di 3,5 parole per enunciato (Fig. 6) mettendo in luce diverse tipologie di struttura. Dove il valore è più basso incide la presenza di enunciati formati da una sola parola, in particolare nei momenti di incertezza e poca spinta comunicativa dei bambini, caratterizzati da frequenti risposte con olofrasi (*sì/no/certo*), soprattutto nel caso di 1M che infatti ha il valore più basso di LME (3,5) con un quarto degli enunciati formati da una sola parola (28/111). Anche 9M (LME: 4,3) si discosta dal valore medio, segnalata in figura dalla "x", restando poco al di sotto della distribuzione maggioritaria del campione. Dato il tipo narrativo-descrittivo di parlato, le sequenze possono allungarsi e farsi più distese e ciò avviene massimamente per 6M (LME: 7,9), outlier della distribuzione.

Figura 6 - *Distribuzione dei valori di Lunghezza Media degli Enunciati*



In relazione alla velocità d'eloquio, il dato di LME evidenzia una corrispondenza diretta fra la variazione dei due parametri (Fig. 7). In altre parole, i bambini che producono enunciati più brevi sono anche quelli con velocità più ridotta. Non segue questo andamento 9M che aggiunge a un andamento prosodico piatto e a una velocità sostenuta la preferenza per enunciati formati da poche parole, uno stile quindi più frammentato.

Figura 7 - *LME e velocità d'eloquio*



Per quanto riguarda il contenuto locutivo, le registrazioni presentano particolari fenomeni nella produzione linguistica, tra cui il ricorrere di frammenti di parola. Oltre alle disfluenze tipiche del parlato spontaneo, sono stati individuati nel campione tre fenomeni rilevanti: a) parola spezzata da confine prosodico; b) accorciamento di parola, in cui può mancare la parte iniziale o quella finale; c) ripetizione di frammento finale di parola. Se ne vedano degli esempi nei seguenti enunciati estratti dal campione:

a) saluta la fami / glia di rane / e torna a casa // [5M]
b) il / bino si è arrabbiato // [1M] (bino per bambino)
c) e il cervo / rvo / corre / verso il cane per fatti propri / con il bambino impiglia-to // [7M]

I primi due tipi di frammentazione sono presenti nel campione per intero; il terzo, ascrivibile all'insieme dei fenomeni di palilalia spesso riscontrati negli individui dello spettro autistico (Karmali, Greer, Nuzzolo-Gomez, Ross, Rivera-Valdes, 2005), risulta specifico di 7M. Molti casi di questa frammentazione atipica si ritrovano in corrispondenza con nessi consonantici. Sono stati inoltre riscontrati in tutti i bambini esempi di parole manipolate (come *is* per *sì*) o inventate (come *sdoprò* per *scoprì*).

4.2.3 Misurazioni di struttura informativa

La segmentazione in enunciati e unità prosodico-informative ha permesso di osservare la distinzione tra enunciati semplici (formati da una sola unità) e complessi (formati da più unità). Il dato si associa a quello di LME nel mostrare il livello di complessità di struttura enunciativa. Si vedano in Tab. 5 le percentuali per partecipante:

Tabella 5 - *Enunciati semplici e complessi*

	Enunciati	*En. semplici %*	*En. complessi %*
1M	111	64,9 %	35,1 %
2M	99	58,6 %	41,4 %
3F	99	51,5 %	48,5 %
4F	144	43,8 %	56,3 %
5M	91	41,8 %	58,2 %
6M	93	24,7 %	75,3 %
7M	103	35,9 %	64,1 %
8M	102	35,3 %	64,7 %
9M	115	50,4 %	49,6 %

I primi tre partecipanti mostrano una maggioranza di strutture informative complesse e articolate in più unità, mentre il resto del campione ha il comportamento opposto; 9M presenta infine un numero quasi pari fra i due tipi di enunciati. Il tipo di discorso, prevalentemente descrittivo-narrativo, può dare spazio al parlante per svolgere turni più lunghi e maggiormente articolati rispetto al veloce botta e risposta di enunciati semplici che solitamente caratterizza i dialoghi; non stupisce quindi la prevalenza, seppur lieve, di enunciati complessi considerando il campione nel suo complesso. Come osservato per il dato di LME, il valore maggiore è quello di 6M, che si mostra particolarmente disinibito nei primi minuti di registrazione, dilungandosi in spiegazioni poco inerenti al contesto nella parte di presentazione personale. Riportiamo a titolo esemplificativo in d) il secondo turno di 6M della durata totale di 01:05 min, in risposta alle iniziali domande per far parlare il bambino di sé[14]:

d) allora / di solito gioco a_ / Minecraft / eh_ / che più o meno lo conoscono tutti // poi Undertale / che è_ / un gioco che lo conoscono_ / un po' in meno / che parla [/2] che c' è questa storia / che prima / c' erano due razze che [/1] che / dominavano la terra // mostri / e umani // [p: 0,45s] poi però / scoppiò una guerra tra queste due razze // e_ gli umani vinsero // [p: 0,35s] e praticamente / in questo gioco / gli umani hanno rinchiuso / i mostri / &s [/1] nel sottosuolo / con una barriera // dalla quale si può &ntrare / ma non si può uscire // [p: 0,75s] e_ / il gioco parla di questo umano caduto / nel sottosuolo / e_ / devi continuare il gioco / e ci sono tre finali // c' è / la_ / pacifist run / che sarebbe / il finale dove diventi amico di tutti / [sospiro] e_ / poi vinci il gioco // poi c' è la neutral / dove [/1] che è un_ / po' più particolare // perché / puoi decidere se / uccidere i mostri / o diventare loro amico // [p: 0,40s] se uccidi anche solo uno di loro / farai una neutral // poi c' è la genocide / che lo dice anche la parola / dove uccidi tutti quanti // e quindi + [p: 0,55s] eh_ / [sospiro] ci sono anche / delle cose / un po' paurose di questo gioco // perché +

[14] La trascrizione riportata utilizza, oltre alla segnalazione dei confini prosodici terminali e non terminali, alcune notazioni particolari quali: "[/n]" per i fenomeni di disfluenza con il numero delle parole che vengono ritrattate; "_" per gli allungamenti vocalici; "&" a inizio di frammento di parola; "+" come segnalazione di enunciato interrotto (in questo caso lasciato in sospeso, non interrotto dall'intervento dell'interlocutore); fra parentesi quadre le durate delle pause presenti fra gli enunciati e i sospiri.

Al contrario, la frammentazione dell'informazione in enunciati semplici aumenta dove già era stato notato un valore basso di LME (1M, 2M, 3F e 9M). Si veda in e) un esempio di scambio tra 2M e l'intervistatrice in cui lo stile telegrafico è evidente così come la difficoltà del bambino nel proseguire il racconto della *Frog Story* che porta l'intervistatrice a inserirsi con suggerimenti:

e) INT: hanno ...
2M: [p: 0,20s] ritrovato la &fami //
INT: [p: 0,35s] la rana //
2M: la rana scomparsa //
INT: [p: 0,35s] però / con chi è la rana ?
2M: [p: 0,40s] con un' altra rana //
INT: [p: 0,20s] con la sua ...
2M: [p: 0,35s] famiglia // [2M]

Dall'analisi delle registrazioni è parso rilevante sottolineare l'uso e la frequenza di segnali discorsivi e riempitivi da parte dei bambini. I più frequenti sono *tipo*, *cioè*, *poi*, usati sia come connettivi scarsamente semantizzati che come veri e propri intercalari e presenti quasi esclusivamente nella parte di narrazione. Un uso tipico del *poi* è quello di marca di inizio enunciato o di cambio di immagine nella *Frog Story*. Ricorrono poi vocalizzazioni prolungate (*eh*/*mh*) come riempitivi. Si vedano gli esempi sottostanti[15]:

f) poi / &co [/1] iniziano [/1] poi il cane / **tipo** annusa delle cose / e vedono che si dirige verso un bosco / e_ / **tipo** vanno verso l' alveare // [8M]

g) **e niente** / sono andati a cercare la rana / [p: 0,65s] e_ / [p: 1,85s] **poi** hanno provato a cercarla / in una tana / soltanto che **poi** in quella tana c' era / [p: 0,50s] un topo // [6M]

La frequenza dei segnali discorsivi e riempitivi è stata calcolata in rapporto alla durata degli enunciati, poi normalizzata a 100s per rendere il parametro confrontabile tra i soggetti (Tab. 6):

Tabella 6 - *Frequenza di segnali discorsivi e riempitivi*

	SD/100 s	*SR/100 s*
1M	0,606	5,456
2M	1,757	0,439
3F	2,687	2,687
4F	5,800	0,773
5M	0,988	1,977
6M	5,104	4,711
7M	1,015	0,000
8M	8,695	7,065
9M	0,675	0,675

[15] Fra parentesi quadre sono segnalate nell'esempio le durate delle pause all'interno dell'enunciato; "..." segnala intonazione volutamente sospensiva.

Dai dati qui riportati spicca l'uso massiccio che 8M fa dei segnali discorsivi e riempitivi, come nell'esempio f), soprattutto nella parte dedicata alla *Frog Story*. A parte 7M e 9M, i più grandi, che quasi non fanno uso di queste forme, segnali discorsivi e riempitivi aiutano i bambini a reperire referenti, far progredire la narrazione e tenere il turno.

5. *Conclusioni*

Sintetizzando i risultati, le analisi acustiche hanno evidenziato un comportamento eterogeneo del campione e messo in luce come fondamentale per la descrizione dell'alterazione prosodica la combinazione range di F0 in semitoni e cvint, al di là quindi dell'altezza tonale dei singoli soggetti. Tramite questi parametri è possibile descrivere la monotonia o al contrario l'enfasi percettivamente riscontrata all'ascolto. Dall'osservazione di tutti i parametri considerati emergono come estremi i comportamenti di due soggetti, 3F e 9M, la cui prosodia risulta quindi più atipica all'interno del campione considerato, esageratamente enfatizzata per 3F e marcatamente monotona per 9M. Si può ipotizzare infine una relazione inversa tra cvint e velocità d'eloquio, per cui a un parlato meno variato in intensità corrisponde nel campione una maggiore velocità. Le misurazioni lessicali e di struttura informativa hanno permesso di osservare invece un differenziarsi di strutture da uno stile olografico e telegrafico a uno a cascata di enunciati complessi, in base alla maggiore o minore responsività e comunicatività dei bambini. Da segnalare, le particolari forme di frammentazione riscontrate come la presenza di parole spezzate da confine prosodico, gli accorciamenti di parola e le ripetizioni palilaliche di frammenti finali di parola.

Le alterazioni prosodiche e di struttura informativa che l'analisi ha messo in luce riflettono l'eterogeneità attesa, trattandosi appunto di Spettro dell'Autismo. I dati qualitativi e quantitativi mostrano inoltre l'estremizzazione di alcuni parametri che possono risultare utili per approfondimenti futuri. Da qui, due filoni principali guidano gli sviluppi della ricerca: da una parte l'ampliamento del campione e il reperimento di dati di confronto di bambini a sviluppo tipico, dall'altra la sensibilizzazione nei confronti della rilevanza della componente prosodica del linguaggio, spesso considerata marginale rispetto alle altre componenti linguistico-comunicative all'interno della ricerca e della pratica logopedica. Attualmente, di fatto, la prosodia non è un'area inclusa nella valutazione e tanto meno nel trattamento logopedico ordinario dei soggetti con ASD; tuttavia, le difficoltà prosodiche sono invalidanti dal punto di vista comunicativo e dell'integrazione sociale, e tendono a permanere anche quando le altre componenti linguistiche migliorano (McCann, Peppé, 2003).

Ringraziamenti

Si ringrazia la Fondazione MAiC di Pistoia, sede di Barile, per la collaborazione e l'accesso a struttura. Un ringraziamento particolare va alla Dott.ssa Notarangelo per il supporto clinico e tecnico alla realizzazione del progetto.

Riferimenti bibliografici

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, DSM-5. Arlington. [It. Transl.: (2014) Manuale diagnostico e statistico dei disturbi mentali, Quinta edizione, DSM-5. Milano: Raffaello Cortina Editore].

Barbosa, P.A. (2019). *Prosódia*. São Paulo: Parábola Editorial.

Barone, R., Bravaccio, C., Muratori, F., Salomone, E., Venuti, P. & Vicari, S. (2020). *Disturbi dello Spettro Autistico*. Milano: EDRA.

Boersma, P., Weenink, D. (2022). Praat: doing phonetics by computer [Computer program], Version 6.2.06. https://www.praat.org

Brambilla, P. (2011). Competenza pragmatica e comprensione delle emozioni Training studies con bambini a sviluppo tipico e atipico. Tesi di Dottorato, Università degli Studi di Milano.

Chiti, V. (2022). L'alterazione della Prosodia nel Disturbo dello Spettro Autistico: analisi del parlato di un gruppo di bambini in età scolare. Tesi di laurea triennale in Logopedia, Università degli Studi di Firenze.

Cresti, E. (2000). Corpus di italiano parlato. In *Studi di grammatica italiana*. Firenze: Accademia della Crusca, I-II.

Cresti, E., Moneglia, M. (1997). L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In Bortolini, U., Pizzuto, E. (a cura di), *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*. Pisa: Edizioni del Cerro, II, 57-90.

Cresti, E., Moneglia, M. (2005). C-ORAL-ROM: Integrated reference corpora for spoken Romance languages. Amsterdam: Benjamins.

Cresti, E., Moneglia, M. (2010). Informational Patterning Theory and the Corpus based description of Spoken language. In Moneglia, M., Panunzi, A. (a cura di), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze: Firenze University Press.

Cresti, E., Panunzi, A. (2013). *Introduzione ai corpora dell'italiano*. Bologna: Il Mulino.

D'Amico, S., Devescovi, A. (a cura di) (2013). *Psicologia dello sviluppo del linguaggio.* Bologna: il Mulino.

Eigsti, I., Schuh, J., Mencl, E., Schultz, R. & Paul, R. (2011). The neural underpinnings of prosody in autism. In *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*.

ISAAC (2017), Documento preliminare ISAAC Italy su CAA e Disturbo dello Spettro dell'Autismo, sito web ISAAC Italy, http://www.isaacitaly.it/wp-content/uploads/2022/11/doc-autismo-e-caa-isaac-italy-2022.pdf

Izre'el, S., Mello, H., Panunzi, A. & Raso, T. (a cura di) (2020). *In Search of Basic Units of Spoken Language: A CorpusDriven Approach*. Amsterdam: Benjamins.

Kanner, L. (1943). Autistic disturbances of affective contact. In *Nervous Child*, 2(3), 217-250.

Karmali, I., Greer, R.D., Nuzzolo-Gomez, R., Ross, D.E. & Rivera-Valdes, C. (2005). Reducing palilalia by presenting tact corrections to young children with autism. In *The Analysis of verbal behavior*, 21(1), 145–153.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Mayer, M. (1969). *Frog, where are you?*. New York: Dial Press.

McAlpine, A., Plexico, L.W., Plumb, A.M. & Cleary, J. (2014). Prosody in Young Verbal Children With Autism Spectrum Disorder. In *Contemporary Issues in Communication Science and Disorders*, 41, 120-132.

McCann, J., Peppé, S. (2003). Prosody in autism spectrum disorders: a critical review. In *International Journal of Language & Communication Disorders*, 38(4), 325-350.

Ministero della Salute (2023). https://www.salute.gov.it/portale/saluteMentale/dettaglioContenutiSaluteMentale.jsp?lingua=italiano&id=5613&area=salute%20mentale&menu=vuoto

Tager-Flusberg, H., Paul, R., Lord, C., Volkmar, F., Paul, R. & Klin, A. (2005). Language and communication in autism. In Volkmar, F., Paul, R., Klin, A. & Cohen, D. (a cura di), *Handbook of Autism and Pervasive Developmental Disorders*. Hoboken, New Jersey: John Wiley & Sons, Inc, (1), 335-364.

Testi, A., Padovani, R., Bilancia, G., Barbi, G. & Musetti, A. (2015). Comprensione delle caratteristiche prosodiche nello sviluppo tipico e in soggetti con Disturbo dello Spettro Autistico. Dati preliminari con prove originali. In *Autismo e disturbi dello sviluppo*, 13(1), 51-64.

Young E.C., Diehl J.J., Morris D., Hyman S.L., Bennetto L., (2005). The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders. In *Language, Speech, and Hearing Services in Schools*, 36(1), 62-72.

PARTE III

VARIAZIONE E APPRENDIMENTO DELLA SECONDA LINGUA

VARIATION AND SECOND LANGUAGE LEARNING

DOI: 10.17469/O2111AISV000015

DUCCIO PICCARDI

Reference, sound symbolism, and variation. In search of an elusive link

This paper deals with an unresolved issue concerning Ohala's frequency code proposal on the origins of sound symbolism. To date, there is a lack of evidence supporting the assumed human capability of extracting meaning features from the indexical realm and repurposing them in sound-symbolic denotation strategies. Here an experimental replication of this mechanism was attempted, built on previous observations that, in the Florentine variety, long-lag voiceless plosives are invested with indexical meanings related to manliness. Through a forced-choice task, 22 local senior high-school students were asked to match nonwords with or without initial long-lag voiceless plosives with male or female specimens of fictitious animal species. Results failed to highlight the expected association between aspirated nonwords and male specimens; however, stimuli read by a male (vs. female) voice were significantly more likely to be matched with male animals, suggesting that sound-symbolic transfers between meaning domains can indeed be triggered under specific circumstances.

Keywords: sound symbolism, denotation, indexicality, frequency code, Voice Onset Time.

1. *Sound symbolism: from maluma to hero*

About a century ago, almost contemporarily, two researchers of very different backgrounds made us aware that, indeed, people actively follow the principle that sounds can bear specific meanings and associations to referential features, which was previously only hinted in anecdotes, theoretical recounts, or through lexical analyses (Magnus, 2013). On the one hand, involving more than 500 participants, anthropologist-linguist Sapir (1929) tested the hypothesis that vowel quality evokes referential size (*mal-mil* effect); on the other, psychologist Köhler (19305: 242-243), elaborating on Usnadze (1924), showed his audience two abstract figures, one being spikey, the other curvilinear, while elusively stating that "if [...] the reader is asked to choose which he would rather call *takete* or *maluma*, he will be probably be able to decide with ease". The success of two major schools viewing the experimental tradition on what we will call from now on "sound symbolism" as outside their primary research interests slowed down its recognition in the mainstream disciplinary scenario (Magnus, 2013): Saussurian structuralism, with its centrality of arbitrariness, and generativism, with its focus on Grammar and its troubled relationship with stochastics and semantics (Kawahara, 2020; see also

Alderete, Kochetov, 2017)[1]. As a consequence, reports have begun to pop up on the quite strict (Svantesson, 2017), quasi-demonizing (Nobile, 2019) marginalization of sound symbolism in the academia.

In recent times, several studies underlined that sound symbolism is currently enjoying a second youth, which is signalled through a steady increase in the related publications (*ibid.*; Sidhu, Pexman, 2018; Nielsen, Dingemanse, 2021). Adopting a rather pragmatic mindset, Shinohara, Uno (2022) tried to interpret this novel contingency. From their viewpoint, sound symbolism succeeded in overcoming disciplinary resistances since it was proven to be a valuable asset in refining our knowledge about crucial research topics, such as the organization of the lexicon, language acquisition, and its origins. These findings, in turn, prompted to relativize "the Saussurian dictum of arbitrariness" *(ibid.*: 2). Nobile (2019) looked at this change of attitude through wider sociological lens. In his recount, Saussurian arbitrariness of the sign is seen as a postulate of demarcation between the human and the natural sciences, or, latu senso, between (rational) culture and (physical) nature. However, in contemporary times, economic and environmental crises induce scientific paradigms to rethink man as essentially bound to the rules of the physical world. This may explain the surge of interest in embodiment and enaction in the cognitive disciplines, and a greater tolerance to sound symbolism (i.e., non-arbitrariness) in linguistics. In this context, it comes as no surprise that researchers are willing to find a legitimate place for sound symbolism in their disciplinary milieu. An extreme case in point is Kawahara's (2020) attempt at promoting it among theoretical phonologists, which provides insights on why researchers trained in the generative tradition should care about this topic. Indeed, recent disciplinary trends in phonology, such as a renewed attention to phonetic events (Hannahs, Bosch, 2018), as well as to gradient and probabilistic phenomena (Alderete, Finley, 2023) seem to create new opportunities for this field to develop an interest in sound symbolism.

In this paper, I will deal with another field which is currently striving to legitimate sound symbolism in its tradition: variationist sociolinguistics (§ 1.1). In particular, by addressing the underspecified relationship between indexicality and denotational meaning in a renowned model on the origins of sound symbolism (the frequency code, § 1.2), I will join the contemporary interest in closing theoretical gaps (e.g., Sidhu, Pexman, 2018; Ekström, 2022). In order to do that, I will propose a forced-choice experiment (§ 4) to 22 Florentine Italian students under the assumption that, since long-lag voiceless plosives are locally invested with social meanings related to manliness (§ 2), they should be able to sound-symbolically extract this feature and

[1] At the risk of oversimplifying, generativism traditionally placed probabilistic tendencies outside linguistic competence, while dichotomous, categorical distinctions (grammatical vs. ungrammatical) are pertinent to grammar (see also Alderete, Finley, 2023). Moreover, the Chomskyan tradition gave a marginal role to semantics (Piattelli-Palmarini, Cecchetto, 1997), starting with the skepticisms of Chomsky himself. See Glanzberg (2021) for a historical overview on semantics in Chomskyan work, especially with respect to the autonomy of syntax, and also Krivochen, Lacková (forthcoming)'s discussion on Chomskyan rejection of externalism as a potential problem for the analysis of iconicity in the generative framework.

coherently use it in denotation. Methodological considerations are expounded in § 3, while the results are discussed in § 5. Lastly, conclusions and directions for future research are drawn in § 6.

1.1 The social life of sound symbolism

The disciplinary intersection between sound symbolism and sociolinguistics is, again, one of initial skepticism and recent enthusiasm. An early systematic attempt at finding a place for it in the Labovian paradigm can be read in Gordon, Heath (1998). The authors argue that the role of speaker sex in language variation and change can be traced back to sound-symbolic preferences based on the dimorphism of the phonatory apparatus (Ohala, 1984; more on this in § 1.2). In other words, women are expected to be leaders in raising/fronting, while men in lowering/backing phenomena. In the comments reported after the main text of the paper, this proposal was deemed as simplistic at best, and its predictions were consequently never systematically tested for (Eckert, 2019: 764). Nonetheless, Labov himself picked up this idea with interest, while ultimately dismissing it because of too many counterexamples (Labov, 2001: 291-292; but see the reply in Heath, 2003) and its determinism (Eckert, Labov, 2017). It should be noted that Gordon and Heath's (1998) theory was published in an anthropology journal. Indeed, at that time, anthropological linguistics had already entertained fruitful discussions on the roles of sound symbolism, especially in the expression of affect (e.g., Samarin, 1978; Ochs, Schieffelin, 1989; Besnier, 1990). These reflections will have a clear echo in the more recent inclusion of sound symbolism in the sociolinguistic mainstream.

Formalized in Eckert (2012), the third wave of variationist sociolinguistics has clear ethnographic origins (*ibid*.: 93). This successful approach shifted the focus of variation to individual stylistic practices, meaning-making processes, and indexical mutability. The study of socio-semiotics broadens the scope of the sociolinguistic inquiry beyond the realm of language per se. Retracing the steps of indexical elaborations, sociolinguistics ends up becoming an embodied discipline concerned with discourse as part of the material world (Bucholtz, Hall, 2016). Mirroring Nobile's (2019) considerations (§ 1), this drift coherently led to a rediscovery of sound symbolism. In Eckert (2010), the author explains the affective use of backed/fronted vowel variants enacted by two preadolescent girls as the result of contextual indexical connections stemming from size sound symbolism, which "lies somewhere between the natural and the conventional" (*ibid.*: 70). Lee (2020) expanded our dictionary of social meanings flourishing from size sound symbolism noting that vowel frontness is perceived as "refined" by speakers of Baba Malay. Similar perspectives are adopted for the explanation of consonant phenomena, such as sibilant variation (Levon, Maegaard & Pharao, 2017). Thus, sound symbolism is now included in seminal overviews as a fundamental indexical resource (e.g., Eckert, 2019; D'Onofrio, Eckert, 2021) and, as Woschitz (2019) argutely observed, its acceptance in sociolinguistic theory signals the emergence of interferences with Labovian standpoints, such as the assumed social meaninglessness of specific

variation (and change) phenomena. Moreover, the fact that individual sound-symbolic resources can point to both the qualities of the denotatum (e.g., [i] means small object) and the speaker alike (e.g., [i] means small utterer) strengthens an emerging sociolinguistic view (e.g., Johnstone, 2010) positing fluid boundaries between the referential and the social meaning domains (D'Onofrio, Eckert, 2021: 30; but see Munson, 2010).

Being framed in the larger picture of iconicity as the ("conventional but [...] invok[ing] the natural") property of a sign whose "vehicle is construed as resembling its object in some diagrammatic or sensual way" (D'Onofrio, Eckert, 2021: 33), sociolinguistic sound symbolism found its ideal phylogenetic framework in Ohala's (1984) "ethological" theory, the frequency code, which is grounded in animal and human physiology. In the next paragraph, I will focus on the frequency code, interpreting it as a theory of indexicality which crucially underspecifies its presumed relationship with denotational meaning.

1.2 Disentangling the frequency code *patola*

In its original formulation, Ohala's (1984) frequency code theory is presented as a Gujarati *patola* fabric, and each of its arguments as a colorful thread which should be evaluated only after being woven into the pre-planned pattern. This rhetorical structure makes the frequency code a "highly synergetic", "synthetic", "bold" proposal (Winter, Oh, Hübscher, Idemaru, Brown, Prieto & Grawunder, 2021: 2) in need of further inquiries.

Ohala posited that physical correlations found in the animal world, e.g., the direct correspondence between the overall size of a living being and the mass of its vibrating membranes ("vocal cords in mammals, syrinx in birds"; Ohala, 1984: 5), and the inverse correspondence between the latter and the height of the frequencies they can produce, originated several linguistic phenomena, including social indexing and size (referential) sound symbolism. By disposing Ohala's threads in a causal order, we can infer that frequency as a feature carrying biologically determined indexicality (i.e., signaling body size; Foulkes, 2010) was involved in basic (socio) semiotic reinterpretations, and thus contextually used to enact hostile stances and exploit the concept of size in interindividual competition. These prompted a sexual selection in the human species, as successful males manifested enlarged phonatory apparatuses capable of enhancing their aggressive displays, and ended up with the establishment of our sexual dimorphism, which in turn added a new layer of sex-based biologically determined indexicality to human voice (i.e., signaling speaker sex through different average F_0). Thus, Ohala's theory concerns the role of indexicality in shaping our reality, retracing several semiotic moves between the natural and the conventional. The other proposed threads concerning these facets of meaning, such as the affective uses of F_0, seemingly match the colors of the general pattern, as they can be interpreted (just as the other sociolinguistic phenomena expounded above, § 1.1) as the continuation of an indexical order (Eckert, 2008 and references therein) stemming from general co-occurrences regarding frequency. On the contrary, the

addition of referential meaning (i.e., tonal and segmental sound-symbolism) to the pattern stands out as a discolored motif. The formation of sound-symbolic mappings onto referential features lies somewhere in the course of this indexical process, but what triggered this osmosis between meaning domains is not entirely clear. Ohala (1984: 5-6) is aware that the explanation of this type of sound symbolism is "more problematic" in his theory, since speakers using frequency resources for the denotation of size are "presumably not trying to appear small or large [...], or even submissive or dominant". The author only briefly addresses this point by stating that frequency serves as a multipurpose trigger of hearer reaction to a feature of the communicative environment, either being part of the utterer or the denotatum.

The vagueness of the frequency code in this regard is baffling if we consider that on the line dividing indexical from referential sound symbolism lie important disciplinary boundaries: in fact, psycholinguists explicitly question the inclusion of the former in their analytical agenda, while still mentioning Ohala's proposal among the main theories covering sound-symbolic associations (e.g., Sidhu, Pexman, 2018: 1626, 1631; Akita, 2021: 3). On the other hand, third-wave sociolinguistics (e.g., Eckert, 2010; Levon, Maegaard & Pharao, 2017) usually relies on the observations advanced in Silverstein (1994) in order to substantiate potential sound-symbolic mergers between meaning domains. Silverstein (1994) reports that Wasco-Wishram has a highly structured variation system based on sound symbolism to express diminutives/augmentatives. What is interesting here is that this system blends into the affective/socio-indexical realm, as the variants convey properties of both the referent and the speaker ("affectively engaging smallness" vs. "affectively engaging bigness"). However, note that diminutives are known to simultaneously express meanings belonging to different domains independently of their means of expression (e.g., Jurafsky, 1996; Dressler, Merlini Barbaresi, 1992 for Italian), so that the sound-symbolic vehicle of Silverstein's examples might be incidental.

Overall, finding ulterior evidence in favor of Ohala's hypothesized evolutionary process may be beneficial for both the utilization of the frequency code in psycholinguistics and the promotion of the transversal acceptation of meaning endorsed by sociolinguistics. Thus, the time seems ripe to move to an experimental approach and test if speakers can transfer meaning from the indexical to the denotational domain. In the next paragraph, I will present the social-meaning-rich feature that will be deployed in this work: Voice Onset Time.

2. *Voice Onset Time circling around hyperarticulation*

Formalized in Lisker, Abramson (1964), Voice Onset Time (VOT) is a sub-segmental feature corresponding to the lapse between the release of a voiceless plosive and the onset of periodicity of the following vowel (Francis, Ciocca & Yu, 2003). Scobbie (2006) strongly promoted the idea that VOT, in addition to its role as a phonological cue, can be invested with indexical values. Indeed, VOT is phonetically enhanced through hyperarticulation (e.g., Schertz, 2013); thus,

long-lag VOT segments represent the fortis endpoint of variation systems based on gradient articulatory strength, and the construction of social meaning on these variants has to take into account this generality (Acton, 2021: 121).

Developing the flagship representational tool of third-wave sociolinguistics, the indexical field, Eckert (2008) gathered together several reports on the social meaning attributed to hyperarticulated /t/ release in American English and disposed them in a descriptive graphical arrangement in which proximity between tags and distance from the center loosely represent plausible socio-semiotic linkages and reinterpretative moves, or increases in indexical order, without predicting fixed directions. Expanding and critically analyzing Eckert's field, Drummond, Schleef (2016: 53) noted that, in order to hold together the disparate meanings of /t/ release, it is necessary to presuppose a central indexical value shared by all the instantiations of the phenomenon, i.e., something related to hyperaticulation ("extremely articulate or clear speech").

The indexical field of hyperarticulated voiceless plosive releases (e.g., Piccardi, 2021: 58) does justice to its primacy in the development of this line of research, showing the tool's potential of providing ideological connections between otherwise irreconcilable meanings (Gafter, 2021). In fact, hyperarticulation originates the impression of both accuracy and emphasis, which can lead to antithetic high-order meanings. For example, while Karthikeyan, Puts, Aung, Link, Rosenfield, Mackiel, Casey, Marks, Cristo, Patel, Santos & Geher (2023) recently noted that, in American English, men producing aspirated /t/s in word-final position (vs. unreleased /t/s) are perceived as more prestigious than physically dominant by other men, Zimman (2023) discussed the attribution of anger to Black speakers' released /t/s.

In Piccardi (2017a), I commented on a VOT pattern found in the speech of 24 Italian Florentine speakers. In this variety, a widespread, highly salient lenition targets post-vocalic plosives (the *gorgia toscana*; e.g., Marotta, 2008); in addition to that, very few studies (e.g., Castellani, 1960) report a specular phenomenon, the *gorgia enfatica*, strengthening the voiceless plosives in initial, post-consonant and gemination contexts and turning their phonologically congruous short-lag releases into relatively longer VOTs (for example, [tʰu] *tu* 'you', [ˈsɛrpʰe] *serpe* 'snake', [ˈbɛkːʰo] *becco* 'beak'). In my sample, male speakers had longer releases in these contexts; moreover, a qualitative perception follow-up noted that participants showed different levels of awareness of the *gorgia enfatica*. Overall, the *gorgia enfatica* could be defined as a male covert-prestige feature, based on the retrieved meanings of iconized articulatory strength, rusticity (stemming from exaggeration of dialectality) and ideological firmness (stemming from both emphasis and precision). Assuming the definition of masculinity as a "bundle of stances" (Kiesling, 2001: 252), the various contextual meanings of the *gorgia enfatica* can be wrapped together in a local gender identity. Moreover, given the coherent production pattern favoring male speakers, it can be assumed that the indexical order of the socio-semiotic reinterpretations of the Florentine feature loops around masculinity

and hyperarticulation, as the former can be both the endpoint and the start (e.g., Eckert, 2008: 471) of ideological[2] linguistic processes.

It could be worth to try and see if Florentine speakers are able to extract the "male" meaning from the *gorgia enfatica* and use it in denotation. This operation has one major caveat: it tries to pattern gender as a social construct with biological sex. As I reviewed earlier (§ 1.2), the frequency code does allow for early "social" exploitation strategies of size-dependent F_0s. Nonetheless, since VOT length has no certain biological-indexical relationship with utterer's sex (Herd, 2020: 597), the differences between our object of analysis and Ohala's frequencies should be taken into account.

3. *Prepping up for the experiment*

For this first experiment on the sound-symbolic link between social and denotational meaning domains, it seems more than fitting to roll back to the basics of this line of research: forced-choice tasks (e.g., Sapir, 1929). This chapter will deal with the building blocks of this type of protocol: the ideation of an effective experimental frame, the construction of nonword stimuli *(§* 3.1), and their modality of presentation *(§* 3.2). Lastly, a brief note will cover the setting of the experiment *(§* 3.3).

3.1 Nonsense animals

Social meaning feeds on communication context (Eckert, 2012): for this reason, studies on this topic make rare use of nonwords (Munson, 2010: 173). Thus, a good balance between stimulus plausibility and distance from real lexical items (to avoid bias) seems a pivotal concern for the experiment attempted here.

For the investigation of the property of sex, two frames are seemingly viable: human and animal names. An increasing body of evidence suggests that anthroponyms manifest sound-symbolic patterns pointing at both physical and personality traits: for example, sharp-sounding names such as "Kirk" are associated with angular faces and harsh personalities (see the literature reviewed in Sidhu, Pexman, 2019). However, the search for sex patterns in proper names stumbled on the realization that these are subject to local norms of social desirability. The local relevance of names associated with strongly characterized individuals and cultural interpretations of specific sound-human attribute linkages may attenuate or completely override expected naming strategies, such as using "small" vowels (e.g., /i e/) for female anthroponyms (*ibid.*; Wang, 2022: 104). The few hints on Florentine masculinity retrieved from the above-described study are too scant to have a clear picture on this issue leading to a reasonable interpretation of the results. Moreover,

[2] Here and below (§ 5) I use this term following the acceptation promoted through the tradition of studies referenced in Eckert (2008). For example, Irvine, Gal (2000: 35) defined the "ideological aspects of linguistic differentiation" as "the ideas with which participants and observers frame their understanding of linguistic varieties and map those understandings onto people, events, and activities that are significant to them"; the same goes for individual variants. See Irvine (2022) for a related overview.

given the diversity of our setting (e.g., Nodari, Piccardi & Calamai, 2023; § 3.3), controlling for the similarity between nonwords and real anthroponyms can be seen as a noticeable hurdle. Similarly, it has been noted that animal (common) names can manifest sound-symbolic designs (Berlin, 2006), and De Carolis, Marsico & Coupé (2017: 25) endorsed their use in related experimental procedures because of the human cognitive sensitivity to animal features (see the references therein and also Barrett, 2005). Relying on an animal frame could seemingly mitigate the two above-discussed issues related to anthroponymy at the cost of creating a conceptual gap between the supposed trigger of the effect (a socio-indexical feature pertaining to human semiotics) and the endpoint of the forced choice. However, if Ohala's (1984) ethological proposal is at work when people evaluate the size of tables (as in Sapir, 1929), this should be a minor theoretical issue to confront.

In order to create Italian nonword stimuli representing animal names while controlling for psycholinguistically relevant parameters, I relied on WORDEN (Origlia, Cangemi & Cutugno, 2015), a Python interface built on the PhonItalia (Goslin, Galluzzi & Romani, 2013) knowledge base. Through WORDEN, I generated ˈCVCV nonwords while critically considering the following aspects.

– Even though the debate on the issue is far from being settled (Shinohara, Uno, 2022), I relied on Kawahara, Shinohara & Kuchimoto (2008) and assumed that the psycholinguistically prominent first segment of the nonwords was the optimal slot to trigger sound-symbolic effects. For this reason, WORDEN was set as to generate forms containing initial voiceless plosives /k t p/ (33, 11 per type), leaving the manipulation of consonant release to a subsequent phase (§ 3.2). 22 distractors with initial non-plosive segments were also analogously generated. The second consonant (whose properties were inserted as statistical control in the analysis; see § 4.4) was free to vary for the sake of enriching the stimulus battery.
– While the first vowel of the nonwords was controlled by setting it to a default /a/, vowel ending represented a major issue in stimulus construction. In fact, not only Italian marks grammatical gender through vowel ending (which may condition the perception of referent's biological sex; cf. Vigliocco, Vinson, Paganelli & Dworzynski, 2005), but previous research (Danesi, 1998) has also shown that Italian speakers assign grammatical gender to nonwords following sound-symbolic associations with relevant features of the proposed referents. Creating nonword stimuli with agrammatical vowel endings (or removing them tout court) would have implied a drastic reduction of their plausibility as Italian lexical items, hindering the access to local socio-indexical knowledge. For this reason, I decided to embrace this presumably strong grammatical bias and tried to add a balanced number of stimuli with the three Italian singular vowel endings (18 /a/, 20 /e/, 17 /o/) while controlling for their effect in the statistical analysis (§ 4.4). Examples of generated nonwords are /ˈkafa/, /ˈpave/, /ˈtatʃo/, /ˈnama/ (distractor).
– Previous research questioned the role of neighbor effects in sound symbolism (e.g., Preziosi, Coane, 2017). Given the "social" premises of the experiment presented here, having estimates of the number of very similar real Italian words

and their frequency is an invaluable asset for the evaluation of potential stimulus idiosyncrasies. Through WORDEN/PhonItalia, both phonological neighborhood density (M 10.74, SD 2.95) and log frequency (M 2.54, SD 0.51) were computed for each nonword and inserted as a control in the statistical analysis (§ 4.4).

3.2 Acoustic (and visual) stimuli

It has been argued (e.g., Bottini, Barilari & Collignon, 2019) that orthographic experience acts alongside acoustic properties in triggering sound-symbolic effects. In this work, an aural presentation of the stimuli was preferred not only for the reduction of unwanted writing bias, but also because aspirated plosives are not conventionally represented in Italian orthographic norms. Thus, in order to have a high degree of control over the experimental manipulation, aural stimuli based on the above-described (§ 3.1) nonword list were created as follows. Given the centrality of the sex parameter in this protocol, and its effects in steering the social evaluation of aural stimuli (e.g., Laur, 2008), two Florentine speakers (a 26-year-old man and a 27-year-old woman) read the nonword list in the anechoic chamber of the Scuola Normale Superiore of Pisa. Four repetitions of the 55 items list were recorded (44.1 kHz, 24 bit): two unsupervised and two guided productions. In the latter case, speakers were asked to "emit a puff of air" after initial voiceless plosives. In accordance with Kim's (1970: 107) "traditional" definition of aspiration, this instruction replicated the perceptual protocol of Piccardi (2017a) through which the social meanings attributed to the Florentine feature were originally retrieved. The production of an aspirated plosive was pinpointed through positive feedback during this training phase. The best repetitions per condition and speaker were selected in terms of overall clarity. The resulting unsupervised, unaspirated set had an overall mean VOT of 23.3 ms (12.33 SD) - [k], M 34.14, SD 11.06; [t], M 19.23, SD 8.73; [p], M 16.54, SD. 8.83. Adapting Nielsen's (2011) manipulation strategy, the bursts and the releases of these plosives were replaced in Praat (Boersma, Weenink, 2023) with the equivalent components of their aspirated counterparts from the guided productions. This aspirated set had an overall mean VOT of 127 ms (445.06 % increase from the unaspirated set overall mean; 23.41 SD) - [k^h], M 130.95, SD 28.1; [t^h], M 123.36, SD 22.96; [p^h], M 126.68, SD 17.28. The full batch of 176 recordings (((33 targets x 2 aspiration conditions) + 22 distractors) x 2 voices) had a mean duration of 499.68 ms (86.96 SD).

Forced-choice sound symbolic association tasks usually rely on visual stimuli to exemplify the experimentally relevant features of the referents (Köhler, 1930[5]). For the sake of providing participants with a visual reminder of the inquired meaning contrast, I prepared a merged picture of a blue Mars (left) and pink Venus symbol (right) on a white background while keeping an eye on the guidelines expounded in Farnand, Fairchild (2014).

3.3 A brief note on the target population

In order to run my pilot test on the sound-symbolic potential of Florentine long-lag voiceless plosives, I asked for the involvement of a senior high school class of a *liceo scientifico* in Scandicci (Florence; see Piccardi, 2022). Indeed, senior highschoolers are already peeking into the adulthood stage of their social lives (Eckert, 1997), ensuring some level of comparability with the participants in Piccardi (2017a).

Two procedures were conducted alongside the main experiment for the sake of knowing the students better while gathering a set of profiling variables for my statistical analysis. Firstly, they were individually interviewed while covering through a semi-structured topic guide several profiling topics (§ 4.4). Then, they filled in the 44-item Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, De Groot, 1990) through a 7-point Likert scale. This tool was designed to quantitatively estimate five factors pertaining to student engagement in school and learning: self-efficacy, intrinsic value, test anxiety, cognitive strategy use, and self-regulation. In my specific context, these factors could help shed some light on the effects of potential individual approaches to the completion of the experimental procedure; moreover, as Nodari (2022) pointed out, attitudes towards school can be relevant predictors of student involvement in local dialectal practices[3]. This version of the questionnaire was adapted and translated through the help of an expert Italian psychologist, and subsequently tested for its reliability (whole questionnaire: $\alpha = 0.89$; all individual factors: $\alpha \geq 0.69$).

4. *The experiment*

4.1 Participants and setting

22 senior year high school students (15 boys, 7 girls, aged 18 to 19) took part to the experiment, which was conducted in their school library in Scandicci (Florence). The room was divided into three separate activity corners for the contemporary deployment of the profiling interviews, the questionnaire completion, and the forced-choice task on sound symbolism.

4.2 Materials

The experiment makes use of the full batch of 176 aural stimuli and the background image described in § 3. These were inserted in a custom Perceval (André, Ghio,

[3] The author conducted fieldwork in two high schools in Lamezia Terme, a municipality situated in the Southern Italian Region of Calabria, in order to provide a sociophonetic account of a Regional Italian feature – allophonic aspiration targeting post-consonant and geminate voiceless plosives. Her student participants were classified in three "attitudes towards school" levels (negative, neutral, and positive), and this variable emerged as pivotal in the explanation of the aspiration patterns. For example, male speakers with negative attitudes produced the longest VOTs in a read speech task; this result was interpreted as the expression of a rough, "anti-school" male persona embracing local linguistic features as opposed to the supralocal norms promoted through formal education (*ibid.*: 231).

Cavé & Teston, 2016) script run through a Windows laptop. The stimuli were heard through a pair of Sennheiser HD 25 Light closed headphones.

4.3 Procedure

One student at a time entered the library every 20 minutes and, after signing an informed consent form, sat with me for the preliminary interview. Then, I led the student to the questionnaire corner and asked him/her to fill in the MSLQ without any time limits, while one of his/her peers took his/her place at the interview corner. Lastly, the student was accompanied to the pc corner and performed the forced-choice sound-symbolism task. Both the MSLQ and the pc experiment were completed without constant supervision. Nonetheless, I was ready to intervene at any sign of perplexity or difficulty.

The Perceval session had the following structure: firstly, instructions were textually presented, framing the experiment into a fictitious discovery of new animal species, and asking participants to match singularly presented aural stimuli with the supposed name of the male or female specimen. No time limit was coded into the experiment, but participants were invited to make their decisions using the left (for male) or right (for female) keyboard arrows as quickly as possible. Then, the 176 stimuli were presented as a sequence of randomized trials with 1.5 secs of inter-stimulus interval (i.e., a silent black screen). During each trial, one stimulus was reproduced together with the Mars/Venus (§ 3.2) image, and participants had to choose between male and female via the assigned keyboard arrows. Both individual choices and response times (RT) were recorded through Perceval. The mean duration of the Perceval sessions was 7.42 minutes (2.11 SD).

4.4 Research hypothesis, pretreatment, and analysis

The Perceval experiment tested for the hypothesis that aspirated nonwords have a higher probability of being associated with male animal specimens than non-aspirated and distractor stimuli. To this aim, 3,872 trials were recorded through the software. Given the unsupervised experimental procedure and the absence of explicit RT limits, an interquartile range criterion of outlier detection was imposed to the response dataset, ultimately pruning 289 observations (RT > 5,263 ms). The results of the remaining 3,583 trials were dummy coded (1 = M, 0 = F) and inserted as response in a generalized linear mixed effect model structure (binomial family) built in R (R Core Team, 2023; lme4 package: Bates, Maechler, Bolcher & Walker, 2015) including Participant and Item as random factors. The main "Condition" predictor (aspirated, filler, unaspirated) was evaluated together, and in interaction with several controls. Below is a comprehensive list of the tested factors, together with basic descriptive information of the variables which were not previously discussed in § 3.

– Stimulus-related: Target consonant (/k/ vs. /t/ vs. /p/ vs. filler); Vowel ending (/a/ vs. /e/ vs. /o/); Second consonant voicing (37 voiced vs. 18 voiceless in the original

nonword list; cf. D'Onofrio, 2014); Voice sex (male vs. female); Phonological neighborhood density; Log phonological neighborhood frequency.
– Participant-related: Sex (male vs. female); Regionality index score[4] (M 2.68, SD 1.82); a qualitatively assigned, two-level categorization of Attitude towards the Florentine dialect (9 positive vs. 13 nonpositive); a mean of the number of qualifications obtained by the student parents (i.e., Parent education: max. 4, min. 1, M 3.09, SD 0.65); Musical proficiency (10 proficient vs. 12 non-proficient; cf. Hallam, 2017); Pet ownership (15 owners vs. 7 non-owners); Number of languages known by the student other than Italian (M 2.27, SD 0.69); MSLQ Self-efficacy (M 4.56, SD 0.66); Intrinsic value (M 4.99, SD 0.86); Test anxiety (M 3.76, SD 1.55); Cognitive strategy use (M 5.25, SD 0.82); Self-regulation (M 4.64, SD 0.92).
– Session-related: RT (pruned dataset M in ms 1,943.66, SD 1,026.44).

Given the number of factors of interest, an incremental approach to model construction was adopted (e.g., Tordini, Galatà, Avesani & Vayra, 2018). Initially, a baseline model containing the main Condition predictor together with Participant and Item as random factors was built; since an evident pattern by Vowel ending was expected (§ 3.1), this control was immediately added to the baseline model as well. Then, the addition of the other factors as individual predictors or in interaction with Condition or the other significant controls was evaluated through χ^2 likelihood ratio tests and comparisons between Akaike and Bayesian Information Criterion scores (see Hay, Foulkes, 2016 for these three methods in binomial mixed effect model selection processes). Individual factors and interactions were retained if their addition granted significant model differences and a lowering of the relative goodness of fit scores.

4.5 Results

Overall, the "male" category was selected in 1,878/3,583 trials (52.41 %). Vowel ending was of paramount importance in determining participant responses (percentages of "male" per category: /a/, 13.32 %; /e/, 51.36 %; /o/, 94.95 %); nonetheless, once morphological gender was controlled for, two other small, but significant effects emerged: Condition (aspirated, 52.05 %; filler, 48.14 %; unaspirated, 55.56 %) and stimulus Voice sex (female, 50.28 %; male, 54.57 %). Moreover, the interaction between Vowel ending and RTs (z-scored through the scale() function out of convergence necessity) was also retained. Tab. 1 shows the final model, and Fig. 1 manifests the directionality of its interaction (sjPlot package: Lüdecke, 2023).

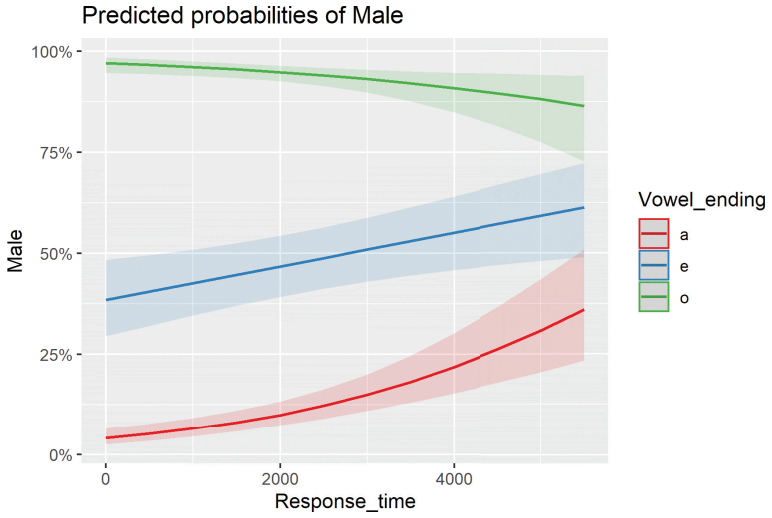
[4] The Regionality Index (Chambers, Heisler, 1999) is an estimate of ethnic affiliation. In the version used in this study, the sum of the scores attributed to the responses to four questions (i.e., student birthplace; student's parent birthplaces; student place of residence; significant relocations) constituted the individual levels of regionality. Each response ranged from 0 ("in Scandicci/Florence") to 2 ("outside the *Provincia di Firenze*") through 1 ("in the *Provincia di Firenze*"), for a maximum summed score of 8.

Table 1 - *Summary of the final model calculating the changes in log odds of responding "male" per unit change of the predictor. Recall that the other categories in the reported variables are Condition (aspirated), Voice sex (female), Vowel ending (/a/). Observations: 3,583. Random factors: Item (variance .08; SD .29); Participant (variance .23; SD .48)*

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.23	0.17	-13.24	< 0.001***
Condition (filler)	-0.34	0.14	-2.51	0.012*
Condition (unaspirated)	0.29	0.12	2.35	0.019*
Voice sex (male)	0.30	0.11	2.85	0.004**
Vowel ending (/e/)	2.09	0.12	16.72	< 0.001***
Vowel ending (/o/)	5.14	0.18	28.41	< 0.001***
z Response time	0.48	0.08	5.91	< 0.001***
z Response time * Vowel ending (/e/)	-0.30	0.10	-3.04	0.002**
z Response time * Vowel ending (/o/)	-0.78	0.14	-5.43	< 0.001***

To sum up, contrary to our expectations, unaspirated nonwords had the relatively highest probabilities of being associated with male specimens, followed by aspirated and filler nonwords (in this order). Moreover, nonwords uttered by a male voice were more likely associated with a male animal specimen than nonwords uttered by a female voice. Lastly, while, overall, participants made their decisions depending on morphological gender, they tended to less morphologically congruent responses as their RT increased.

Figure 1 - *Response Time * Vowel ending interaction plot from the model shown in Tab. 1. RTs were transformed back to ms for the sake of interpretability*



5. *Discussion*

This work tried to provide a first experimental replication of an underdiscussed mechanism pertaining to Ohala's frequency code theory and to its proposed juncture between indexical and referential meaning. Specifically, I posited that Florentine speakers could be able to associate aurally presented nonword stimuli containing socially meaningful long-lag voiceless plosives with referents characterized through properties coherent to the indexical content of the variants. I expected to find an association pattern matching "manly" long VOTs with the names of fictitious male animal specimens. For the sake of guaranteeing the access to the social meaning knowledge of my participants, I decided to build my stimulus battery respecting the structure of real Italian words and included vowel endings marking grammatical gender. This operative necessity provided participants with a clear association-making strategy: responses were mostly guided by vowel ending (i.e., /o/ for "male", /a/ for "female" and /e/ for both), and any other choice pattern was rather weak and dependent on the statistical control of this stimulus feature. Nonetheless, in this chapter, I will advance an interpretation of the present results, aiming to encourage future verifications through similar endeavors.

While the null hypothesis that there is no relationship between response and experimental Condition was indeed rejected through regression modeling, effect directionality run in the opposite direction to the one which was initially advanced: aspirated nonwords had a (slightly) lower, and not higher, probability of being associated with male specimens than their unaspirated counterparts. Moreover, distractors without initial plosive segments had the lowest probability of being associated with male specimens. This result is in line with a strong cross-cultural anthroponymic trend favoring sonorants in female names (e.g., Kawahara, Katsuda & Kumagai, 2019; Wang, 2022), which may depend on associations between femininity and curviness (i.e., *maluma*-like sounds; Köhler, 1930[5]) "dating back millennia" (Sidhu, Pexman, 2015: 17). Marketing research added another piece to our explanatory attempt noting that products with names containing fricatives (vs. plosives) are perceived as more feminine (Klink, 2000). Studies searching for sound-symbolic relationships between plosive VOT and sex are rare. After a corpus analysis of Korean anthroponyms, Sullivan (2020) proposed to both English and Korean participants a name gendering experiment in which pseudo-proper names were evaluated on a scale of appropriateness from female to male. Both plosive VOTs and F_0s of the respective syllable onsets were manipulated. Long VOTs skewed Korean responses towards "male", while this feature had no effect on English associations, leading Sullivan (2020: 58) to conclude that the VOT-sex pattern is language specific. Other than on stark differences between the phonological organizations of the plosive series in the two tested languages[5], the

[5] An anonymous reviewer argued that the weak results of the Italian experiment presented here might heavily depend on phonological factors: Italian does not make distinctive use of aspiration, and aspirated voiceless plosives are absent from its phonological inventory. Indeed, Akita (2015) stressed the

Korean pattern might depend on sound-symbolic consonant alternations evoking emphasis through (phonological) aspirated plosives in word connotation (Kim, 1977), and on plausible ideological links between masculinity and exaggeration (Piccardi, 2017a: 98 and references therein). An alternative explanation could draw Sullivan's results closer to our object of inquiry. Note that her aural stimuli were built on the recordings of a middle-aged Seoul Korean man. In this variety, a sound change in progress is shifting the phonological marking of plosive aspiration from VOT to F_0, with men lagging behind women and thus showing longer release phases. Crucially, it has been argued that this social pattern is perceptually salient (Bang, Sonderegger, Kang, Clayards & Yoon, 2018: 121 and references therein). While any consideration on the perceptiveness of Sullivan's Korean participants to Seoul sociolinguistically grounded VOT distribution should be advanced with caution, voice sex may stand as an important factor in the interpretation of the Florentine results. In fact, the previous comments on the local social meaning of the inquired feature were triggered through a male voice enactment of words containing long-lag VOTs or through participant guided productions (Piccardi, 2017a), while the effects of hearing Florentine female voices using aspiration are currently unknown. As a product of hyperarticulation, long VOTs could be associated with more general female "clear" speech styles, whose frequent manifestations are explained either through sociophonetic or biological accounts (e.g., Munson, 2007: 140 and references therein). Of course, such differential triggering of social evaluations should lead not just to the general association (i.e., with the "female" category, through widespread clear speech patterns) relatively winning over the local one (i.e., with the "male" category, through Florentine masculine speech styles), but to the emergence of a statistically significant interaction between Condition and Voice sex. As this was not the case (Tab. 1)[6], further inquiries on the sound-symbolic consequences of VOT/voice sex social meaning compositionality are warranted.

Moving on to the significant controls, we find that, indeed, Voice sex had a small effect on participant's choices: nonwords uttered by a male voice were more frequently associated with male animal specimens. This result most closely comments on our research interest. Since male mammals are usually bigger than females (e.g., Lindenfors, Gittleman & Jones, 2007), and mammals are "prototypical" animals

idea that phonological typology shapes the emergence of sound-symbolic associations, and Godoy, de Souza Filho, de Souza, França & Kawahara (2020: 719) provided a short review of the literature supporting this view. The monolingual design of the present study is not suited to comment on this intriguing hypothesis, prompting further testing with other languages (§ 6).

[6] In order to double-check this point, I re-added the interaction of interest to the final model shown in Tab. 1. None of the three criteria of variable retention mentioned in § 4.4 favored the more complex model: the $\chi 2$ likelihood ratio test between the Tab. 1 and the Tab. 1 plus the interaction between Condition and Voice sex models did not reach statistical significance, and the latter showed higher AIC and BIC scores, with a difference of more than 2/less than 4 and more than 10, respectively (Raftery, 1995; Burnham, Anderson, 2004). When computed in addition to the Tab. 1 formula, the interaction had the following output: Condition (filler) * Voice sex (male): est. -.2, std. error .27, p = .46; Condition (unaspirated) * Voice sex (male): est. .04, std. error .25, p = .87.

from a student viewpoint (Trowbridge, Mintzes, 1988), the results pertaining to Voice sex may be considered as a roundabout replication of Ohala's F_0-to-size mapping: the low-F_0 human male voice could have been associated with implicitly mammal, generally bigger male animal specimens. However, previous research (Rojczyk, 2011) failed to induce referent size sound-symbolic associations through F_0 manipulation alone. Thus, an alternative explanation of our Voice sex effect could imply a more direct sound-symbolic mapping involving the second frequency code layer of biologically determined indexicality, i.e., the one concerning human sexual dimorphism (§ 1.2). In other words, participants could have extracted general features (including different mean F_0s) of human male vs. female voice and repurposed them to identify living beings in a sexually coherent fashion. While the exact triggers (i.e., the co-variating elements characterizing gendered voice, which may also go beyond the biological into the social realm) of this matching strategy cannot be ascertained through the current experiment, this result highlights that, indeed, speakers can bring out indexical information from acoustic stimuli and predictably use it to depict features of fictitious referents.

Taken together, these two results highlight the role played by co-occurrence consistency (i.e., frequency) between specific indexical meanings and their acoustic means of expression in order to successfully prompt coherent transfers to denotational sound-symbolic mechanisms: the retrieved patterns relatively favored basic socio-semiotic gender characterization of hyperarticulated speech over local social meaning, and general properties of human voice sex. Statistical co-occurrence is indeed one of the main proposed explanatory factors for sound symbolism. Sidhu, Pexman (2018: 1626-1627, 1631) noted that frequency-to-size mappings can be easily framed within this principle, which is also the basis of Ohala's (1984) frequency code. In Piccardi (2017b) I argued that, seeking for a model grounded in the encoding of both eventive frequencies and indexical properties, sound symbolic theory could benefit from the adoption of an exemplarist mindset. Exemplar theory posits that humans are capable of storing memories of individual communicative events in probabilistic structures, retaining fine-grained phonetic details and contextual information (e.g., Pierrehumbert, 2006). In this framework, previous research suggested that contextually evoking specific concepts can activate related language exemplars which, in turn, can steer perceptions (Hurring, Hay, Drager, Podlubny, Manhire & Ellis, 2022) and productions (Sanchez, Hay & Nilson, 2015) towards conceptually coherent targets. Crucially, Hurring *et al.* (2022) recently provided tentative confirmation to the idea that, in order to take place, these contextual priming effects may need a consistent prior exposure of the subject to the linguistic elements linked to the evoked concept. Given this theoretical background, Ohala's (1984) explanation of indexical-to-denotational meaning transfers (§ 1.2) can be rephrased as follows. The biological motivation of F_0 variation in the animal world creates a highly populated cluster of eventive traces encoding information about both vocalization frequency and contextual size (of the utterer). Encountering visual cues denoted by a size-related concept (i.e., "bigness")

activates this clustering, together with its attached acoustic detail. Note that, in the absence of arbitrary word-to-meaning associations (which are roughly minimized using nonwords in sound-symbolic experiments), the speaker's reference exemplar distribution is vastly based on prior indexical experiences. Thus, when prompted to perform a naming (or forced-choice) task, the speaker may be induced to resort to the main acoustic information attached to the activated concept of size, which is originally of indexical nature. While previous research seldom felt the need of making sound symbolism talk to exemplar theory (but see Auracher, Scharinger & Menninghaus, 2019: 12), it may be argued that this framework could lead to a more organic treatment of indexicality and statistical co-occurrence in this line of work[7].

Last in proximity to our research interest but first in magnitude, the effects of Vowel ending on the forced choices provide confirmation to the idea that Italian grammatical gender can affect word meaning, at least when referring to sexed entities (Vigliocco *et al.*, 2005). While a full recount on this topic is well outside the scope of this paper (see Samuel, Cole & Eacott, 2019 for a recent review), it should be noted that /e/ endings, which are "opaque" from the viewpoint of grammatical gender, triggered the most uncertain association pattern and the slowest responses[8] (see the literature gathered in Padovani, Calandra-Buonaura, Cacciari, Benuzzi & Nichelli, 2005: 301), supporting the primacy of grammatical markers in the observed decision process. The interaction between Vowel ending and RT may imply that longer evidence accumulation processes (e.g., Ratcliff, Smith, Brown & McKoon, 2016) are required to reach morphologically unintuitive decisions. However, this result cannot provide any further detail on the type of knowledge added to grammatical gender in the slow responses, nor on its exact source.

6. *Conclusions*

This work finds its place in an ongoing revival of research on sound symbolism. In this milieu, third-wave sociolinguists are enthusiastically considering sound symbolism as an important indexical resource through an endorsement of Ohala's (1984) frequency code, legitimating its presence in their disciplinary paradigm and an incipient acceptation of meaning which encompasses the social and the

[7] This is not by any means to say that exemplar theory is the only model suited to deal with sound symbolism. For example, Kawahara (2021) recently suggested that Maximum Entropy Harmonic Grammar can account for sound-symbolic patterns. Moreover, indeed, other theoretical perspectives in phonology have frequency and variation at heart (Coetzee, Pater, 2011; Alderete, Finley, 2023). Here, exemplar theory is mentioned in reference to the specific issue at hand, which concerns indexical knowledge and might be dependent on the relative frequency of events in which both acoustic and contextual features are concurrently relevant. As mentioned above, the proposal of a probabilistically structured "unified mental architecture" (Pierrehumbert, 2006: 520) embedding these heterogeneous aspects of communicative experience is central to exemplar theory in linguistics (see also Foulkes, Docherty, 2006).

[8] This was verified by fitting a restricted maximum likelihood linear mixed effect model with ms RT as dependent variable, Vowel ending as predictor and Participant and Item as random effects. /e/, and not /o/, (from a baseline /a/) was significant (est. 406.82, std. error 40.33, $p < .001$).

denotational alike (D'Onofrio, Eckert, 2021). Nonetheless, as per admission of Ohala (1984) himself, the frequency code is not well-equipped to justify potential sound-symbolic meaning leakages between these two domains. Thus, I attempted to experimentally verify the supposed human capability of extracting meaning features at the indexical level and repurposing them in denotation by submitting a forced-choice task to a senior high school class in Scandicci (Florence).

The experiment only managed to glimpse a very feeble light at the end of this disciplinary tunnel. For the sake of providing naturalistic nonwords to the social evaluation of Italian participants, I committed to the creation of stimuli with vowel endings marking grammatical gender. This choice had a huge impact on the experimental results, as participants made evident use of them in their decision processes. While, of course, the removal of such bias does not guarantee the emergence of alternative patterns (or the strengthening of the ones commented here), future studies should be primarily concerned with testing Ohala's assumptions in a language without noun gender markers steering participant associations. English is the obvious solution to this conundrum, and not only because of its lack of noun grammatical gender: this language is the homeland of third-wave sociolinguistics and indexical field representations (Eckert, 2008; see § 2), so that similar research can be supported by solid acquisitions on the social meaning nuances of linguistic variables potentially orienting denotational sound symbolism.

Other than vowel ending, two other factors had a very small impact on the experimental results. Aspiration did condition participant reactions, but in an unexpected way: in fact, long-lag nonwords had a little higher chance of being associated with female, and not male, animal specimens. This result was tentatively interpreted as a consequence of relatively prevailing, more general (and thus, less locally bound), patterns between female speakers and clear, hyperarticulated speech, which were probably evoked through the presence of a female voice producing long VOTs in the stimulus battery. Perceived voice sex manifested a less ambiguous pattern: nonwords uttered by a male voice were more likely to be associated with a male specimen. This specific effect gives some plausibility to the idea that, at some level, participants were able to extract the indexical property of sex from the acoustic stimuli and recycle it in the denotation of unknown living entities.

These two almost negligible effects served to highlight a current partial limitation of sound-symbolic research. In the mid-seventies, Hecht (1976) rightly promoted a more articulate modelling of respondent social characteristics in sound symbolism research. Now, sociolinguistics (or sociophonetics) may be called to answer to novel variationist issues: how does frequency of co-occurrence between the two main components of a sound symbolic association (e.g., a sound and an indexical meaning) affect its probabilistic manifestations (cf. Sidhu, Pexman, 2018: 1627)? How can we account for the potential effects of experiment contextual characteristics (e.g., the features of the stimulus utterers, such as sex, age, speech variety, but also environmental conditioning) in protocols investigating sound symbolism? The adoption of

frequency-grounded, contextually informed exemplar models might be the next step for the consolidation of the sociolinguistic tradition on the matter at hand.

Acknowledgements

This paper is the result of a Doctoral spin-off project held at the University of Pisa. I wish to thank my Italian advisors, Giovanna Marotta and Silvia Calamai, for their invaluable guidance. My deepest gratitude goes also to the involved personnel of the Russel Newton high school in Scandicci, namely Anna Maria Addabbo, Susanna Bino, and Luca Gargiulo, as well as to the V C students.

References

Acton, E.K. (2021). Pragmatics and the Third wave: The Social Meaning of Definites. In Hall-Lew, L., Moore, E. & Podesva, R.J. (a cura di), *Social Meaning and Linguistic Variation. Theorizing the Third Wave*. Cambridge: Cambridge University Press, 105-126.

Akita, K. (2015). Sound Symbolism. In Östman, J.-O., Verschueren, J. (a cura di), *Handbook of Pragmatics. 2015 Installment*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/hop.19.sou1

Akita, K. (2021). Phonation Types Matter in Sound Symbolism. In *Cognitive Science*, 45(5), 1-14. https://doi.org/10.1111/cogs.12982

Alderete, J., Finley, S. (2023). Probabilistic Phonology: A Review of Theoretical Perspectives, Applications, and Problems. In *Language and Linguistics*, 24(4), 565-610.

Alderete, J., Kochetov, A. (2017). Integrating Sound Symbolism with Core Grammar: The Case of Expressive Palatalization. In *Language*, 93(4), 731-766.

André, C., Ghio, A., Cavé, C. & Teston, B. (2016). PERCEVAL: PERCeption EVALuation Auditive & Visuelle. [Computer program] Version 4.0.0. https://lpldev.lpl-aix.fr/perceval/

Auracher, J., Scharinger, M. & Menninghaus, W. (2019). Contiguity-Based Sound Iconicity: The Meaning of Words Resonates with Phonetic Properties of their Immediate Verbal Contexts. In *PLoS ONE*, 14(5), 1-21. https://doi.org/10.1371/journal.pone.0216930

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2018). The Emergence, Progress, and Impact of Sound Change in Progress in Seoul Korean: Implications for Mechanisms of Tonogenesis. In *Journal of Phonetics*, 66, 120-144.

Barrett, H.C. (2005). Cognitive Development and the Understanding of Animal Behavior. In Ellis, B.J., Bjorklund, D.F. (Eds.), *Origins of the Social Mind. Evolutionary Psychology and Child Development*. New York / London: The Guildford Press, 438-467.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67(1), 1-48.

Berlin, B. (2006). The First Congress of Ethnozoological Nomenclature. In *The Journal of the Royal Anthropological Institution*, 12(s1), 23-44.

Besnier, N. (1990). Language and Affect. In *Annual Review of Anthropology*, 19, 419-451.

Boersma, P., Weenink, D. (2023). PRAAT: Doing Phonetics by Computer. [Computer program] Version 6.3.10. https://www.praat.org/

Bottini, R., Barilari, M. & Collignon, O. (2019). Sound Symbolism in Sighted and Blind. The Role of Vision and Orthography in Sound-Shape Correspondences. In *Cognition*, 185, 62-70.

Bucholtz, M., Hall, K. (2016). Embodied Sociolinguistics. In Coupland, N. (a cura di), *Sociolinguistics: Theoretical Debates*. Cambridge: Cambridge University Press, 173-198.

Burnham, K.P., Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. In *Sociological Methods & Research*, 33(2), 261-304.

Castellani, A. (1960). Precisazioni sulla gorgia toscana. In *Boletim de Filología*, 19, 242-262.

Chambers, J.K., Heisler, T. (1999). Dialect Topography of Quebéc City English. In *Canadian Journal of Linguistics*, 44(1), 23-48.

Coetzee, A.W., Pater, J. (2011). The Place of Variation in Phonological Theory. In Goldsmith, J.A., Riggle, J. & Yu, A.C.L. (Eds.), *The Handbook of Phonological Theory*. Second Edition. Oxford: Blackwell-Wiley, 401-431.

D'Onofrio, A. (2014). Phonetic Detail and Dimensionality in Sound-shape Correspondences: Refining the Bouba-Kiki Paradigm. In *Language and Speech*, 57(3), 367-393.

D'Onofrio, A., Eckert, P. (2021). Affect and Iconicity in Phonological Variation. In *Language in Society*, 50(1), 29-51.

Danesi, M. (1998). Gender Assignment, Markedness, and Indexicality: Results of a Pilot Study. In *Semiotica*, 121(3-4), 213-240.

De Carolis, L., Marsico, E. & Coupé, C. (2017). Evolutionary Roots of Sound Symbolism. Association Tasks of Animal Properties with Phonetic Features. In *Language & Communication*, 54, 21-35.

Dressler, W.U., Merlini Barbaresi, L. (1992). Italian Diminutives as Non-Prototypical Word Formation. In Tonelli, L., Dressler, W.U. (a cura di), *Natural Morphology-Perspectives for the Nineties*. Padua: Unipress, 21-30.

Drummond, R., Schleef, E. (2016). Identity in Variationist Sociolinguistics. In Preece, S. (Ed.), *The Routledge Handbook of Language and Identity*. London: Routledge, 50-65.

Eckert, P. (1997). Age as a Sociolinguistic Variable. In Coulmas, F. (a cura di), *The Handbook of Sociolinguistics*. Oxford/Cambridge: Blackwell, 151-167.

Eckert, P. (2008). Variation and the Indexical Field. In *Journal of Sociolinguistics*, 12(4), 453-476.

Eckert, P. (2010). Affect, Sound Symbolism, and Variation. In *University of Pennsylvania Working Papers in Linguistics*, 15(2), 70-80.

Eckert, P. (2012). Three Waves of Variation Study. In *Annual Review of Anthropology*, 41, 87-100.

Eckert, P. (2019). The Limits of Meaning: Social Indexicality, Variation, and the Cline of Interiority. In *Language*, 95(4), 751-776.

Eckert, P., Labov, W. (2017). Phonetics, Phonology and Social Meaning. In *Journal of Sociolinguistics*, 21(4), 467-496.

Ekström, A.G. (2022). What's Next for Size-Sound Symbolism? In *Frontiers in Language Sciences*, 1, 1-7. https://doi.org/10.3389/flang.2022.1046637

Farnand, S.P., Fairchild, M.D. (2014). Designing Pictorial Stimuli for Perceptual Experiments. In *Applied Optics*, 53(3), 72-82.

Foulkes, P. (2010). Exploring Social-Indexical Knowledge: A Long Past but a Short History. In *Laboratory Phonology*, 1(1), 5-39.

Foulkes, P., Docherty, G. (2006). The Social Life of Phonetics and Phonology. In *Journal of Phonetics*, 34(4), 409-438.

Francis, A.L., Ciocca, V. & Yu, J.M.C. (2003). Accuracy and Variability of Acoustic Measures of Voicing Onset. In *The Journal of the Acoustical Society of America*, 113(2), 1025-1032.

Gafter, R.J. (2021). Reconciling Seemingly Conflicting Social Meanings. In Hall-Lew, L., Moore, E. & Podesva, R.J. (a cura di), *Social Meaning and Linguistic Variation. Theorizing the Third Wave*. Cambridge: Cambridge University Press, 222-242.

Glanzberg, M. (2021). Chomsky on Semantics. In Allott, N., Lohndal, T. & Rey, G. (Eds.), *A Companion to Chomsky*. Hoboken: Wiley-Blackwell, 416-432.

Godoy, M.C., de Souza Filho, N.S., de Souza, J.G., França, H.A. & Kawahara, S. (2020). Gotta Name'Em All: An Experimental Study on the Sound Symbolism of Pokémon Names in Brazilian Portuguese. *In Journal of Psycholinguistic Research*, 49(5), 717-740.

Gordon, M., Heath, J. (1998), Sex, Sound Symbolism, and Sociolinguistics. In *Current Anthropology*, 39(4), 421-449.

Goslin, J., Galluzzi, C. & Romani, C. (2013). Phonitalia: A Phonological Lexicon for Italian. In *Behavior Research Methods*, 46(3), 872-886.

Hallam, S. (2017). The Impact of Making Music on Aural Perception and Language Skills: A Research Synthesis. In *London Review of Education*, 15(3), 388-406.

Hannahs, S.J., Bosch, A.R.K. (2018). The Study of Phonology in the 21st Century. Overview and Introduction to The Routledge Handbook of Phonological Theory. In Hannahs, S.J., Bosch, A.R.K. (a cura di), *The Routledge Handbook of Phonological Theory*. New York: Routledge, 1-10.

Hay, J., Foulkes, P. (2016). The Evolution of Medial /t/ Over Real and Remembered Time. In *Language*, 92(2), 298-330.

Heath, J. (2003). Review to Principles of Linguistic Change. Volume 2: Social Factors by William Labov. In *Anthropological Linguistics*, 45(4), 466-472.

Hecht, M.L. (1976). More Semantics of Repetition: A Sociolinguistic Investigation of Phonetic Symbolism. In *Communication Quarterly*, 24(3), 33-40.

Herd, W. (2020). Sociophonetic Voice Onset Time Variation in Mississippi English. In *The Journal of the Acoustical Society of America*, 147(1), 596-605.

Hurring, G., Hay, J., Drager, K., Podlubny, R., Manhire, L. & Ellis, A. (2022). Social Priming in Speech Perception: Revisiting Kangaroo/Kiwi Priming in New Zealand English. In *Brain Sciences*, 12, 1-29. https://doi.org/10.3390/brainsci12060684

Irvine, J.T. (2022). Revisiting Theory and Method in Language Ideology Research. In *Linguistic Anthropology*, 32(1), 222-236.

Irvine, J.T., Gal, S. (2000). Language Ideology and Linguistic Differentiation. In Kroskrity, P. (Ed.), *Regimes of Language: Ideologies, Polities, and Identities*. Santa Fe: School of American Research Press, 35-83.

Johnstone, B. (2010). Locating Language in Identity. In Llamas, C., Watt, D. (a cura di), *Language and Identities*. Edinburgh: University Press, 29-36.

Jurafsky, D. (1996). Universal Tendencies in the Semantics of the Diminutive. In *Language*, 72(3), 533-578.

Karthikeyan, S., Puts, D.A., Aung, T., Link, J.K., Rosenfield, K., Mackiel, A., Casey, A., Marks, K., Cristo, M., Patel, J., Santos, A. & Geher, G. (2023). Articulatory Effects on Perceptions of Men's Status and Attractiveness. In *Scientific Reports*, 13, 1-11. https://doi.org/10.1038/s41598-023-29173-z

Kawahara, S. (2020). Sound Symbolism and Theoretical Phonology. In *Language and Linguistic Compass*, 14(8), 1-17. https://doi.org/10.1111/lnc3.12372

Kawahara, S. (2021). Testing MaxEnt with Sound Symbolism: A Stripy Wug-Shaped Curve in Japanese Pokémon Names. In *Language*, 97(4), 341-359.

Kawahara, S., Katsuda, H. & Kumagai, G. (2019). Accounting for the Stochastic Nature of Sound Symbolism using Maximum Entropy Model. In *Open Linguistics*, 5(1), 109-120.

Kawahara, S., Shinohara, K. & Uchimoto, Y. (2008). A Positional Effect in Sound Symbolism: An Experimental Study. In *Proceedings of the Japan Cognitive Linguistics Association*, 8, 417-427.

Kiesling, S.F. (2001). "Now I Gotta Watch What I Say": Shifting Constructions of Masculinity in Discourse. In *Journal of Linguistic Anthropology*, 11(2), 250-273.

Kim, C. (1970). A Theory of Aspiration. In *Phonetica*, 21(2), 107-116.

Kim, K.-O. (1977). Sound Symbolism in Korean. In *Journal of Linguistics*, 13(1), 67-75.

Klink, R. (2000). Creating Brand Names With Meaning: The Use of Sound Symbolism. In *Marketing Letters*, 11(1), 5-20.

Köhler, W. (1930[5]). *Gestalt Psychology*. New York: Liveright.

Krivochen, D.G., Lacková, L. (forthcoming). Iconicity and Generative Grammar. Manuscript, University of Oxford / Palacký University.

Labov, W. (2001). *Principles of Linguistic Change. Social Factors*. Malden/Oxford: Blackwell.

Laur, E. (2008). *Contribution à l'étude des perceptions linguistiques. La méthodologie des faux-couples revisitée*. Montréal: Office québécois de la langue française.

Lee, N.H. (2020). Style Variation in the Second Formant. What Does it Mean to be "Refined" in Baba Malay? In *Language Ecology*, 4(1), 115-130.

Levon, E., Maegaard, M. & Pharao, N. (2017). Introduction: Tracing the Origin of /s/ Variation. In *Linguistics*, 55(5), 979-992.

Lindenfors, P., Gittleman, J.L. & Jones, K.E. (2007). Sexual Size Dimorphism in Mammals. In Blackenhorn, W.V., Fairbairn, D.J. & Székely, T. (a cura di), *Sex, Size and Gender Roles: Evolutionary Studies of Sexual Size Dimorphism*. Oxford: Oxford University Press, 280-315.

Lisker, L., Abramson, A.S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. In *Word*, 20(3), 527-565.

Lüdecke, D. (2023). sjPlot: Data Visualization for Statistics in Social Science. [R package] version 2.8.14, retrieved May, 20 2023 from https://CRAN.R-project.org/package=sjPlot

Magnus, M. (2013). A History of Sound Symbolism. In Allan, K. (a cura di), *The Oxford Handbook of the History of Linguistics*. Oxford: Oxford University Press, 191-208.

Marotta, G. (2008). Lenition in Tuscan Italian (Gorgia Toscana). In Brandão de Carvalho, J., Scheer, T. & Ségéral, P. (a cura di), *Lenition and Fortition*, Berlin: Mouton de Gruyter, 235-271.

Munson, B. (2007). The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation. In *Language and Speech*, 50(1), 125-142.

Munson, B. (2010). Levels of Phonological Abstraction and Knowledge of Socially Motivated Speech-Sound Variation: A Review, a Proposal, and a Commentary on the Papers by Clopper, Pierrehumbert, and Tamati; Drager; Foulkes; Mack; and Smith, Hall, and Munson. In *Laboratory Phonology*, 1(1), 157-177.

Nielsen, A.K., Dingemanse, M. (2021). Iconicity in Word Learning and Beyond: A Critical Review. In *Language and Speech*, 64(1), 52-72.

Nielsen, K. (2011). Specificity and Abstractness of VOT Imitation. In *Journal of Phonetics*, 39(2), 132-142.

Nobile, L. (2019). Introduction. Sound Symbolism in the Age of Digital Orality. A Perspective on Language Beyond 'Nature' and 'Culture'. In *Significances (Signifying)*, 3(1), XXXVI-LXVIII.

Nodari, R. (2022). *L'identità linguistica regionale degli adolescenti. Aspirazione delle occlusive sorde in Calabria e percezione della varietà locale*. Roma: Aracne.

Nodari, R., Piccardi, D. & Calamai, S. (2023). Lingue conosciute, parlate, percepite. Il punto di vista degli studenti in Toscana. In *EL.LE*, 12(3), 385-431.

Ochs, E., Schieffelin, B. (1989). Language has a Heart. In *Text*, 9(1), 7-25.

Ohala, J.J. (1984). An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. In *Phonetica*, 41(1), 1-16.

Origlia, A., Cangemi, F. & Cutugno, F. (2015). WORDEN: a PYTHON Interface to Automatic (Non)word Generation. In Vayra, M., Avesani, C. & Tamburini, F. (a cura di), *Language Acquisition and Language Loss. Acquisition, Change and Disorders of the Language Sound Structure*. *Studi AISV 1*. Milan: Officinaventuno, 459-470.

Padovani, R., Calandra-Buonaura, G., Cacciari, C., Benuzzi, F. & Nichelli, P. (2005). Grammatical Gender in the Brain: Evidence from an fMRI Study on Italian. In *Brain Research Bulletin*, 65(4), 301-308.

Piattelli-Palmarini, M., Cecchetto, C. (1997). The Problem of Meaning in Generative Grammar. In Mandell, C., McCabe, A. (a cura di), *The Problem of Meaning: Behavioral and Cognitive Perspectives*. Amsterdam: Elsevier, 415-469.

Piccardi, D. (2017a). Sociophonetic Factors of Speakers' Sex Differences in Voice Onset Time: A Florentine Case Study. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (a cura di), *Social and Biological Factors in Speech Variation. Studi AISV 3*. Milan: Officinaventuno, 83-106.

Piccardi, D. (2017b). The Social Life of Sound Symbolism. Analysing the Relationship Between Indexical and Referential Meaning. Paper presented at CODISCO X, September 29, 2017.

Piccardi, D. (2021). *Rural Sociophonetics. Dialectology Meets Usage-based Theory in the Village of Antona (Massa, Italy)*. Alessandria: Edizioni dell'Orso.

Piccardi, D. (2022). Teaching Sound Symbolism through Japanese Pop Culture. A Response to Kawahara (2018). In *Italiano LinguaDue*, 14(2), 724-758.

Pierrehumbert, J.B. (2006). The Next Toolkit. In *Journal of Phonetics*, 34(4), 516-530.

Pintrich, P.R., De Groot, E.V. (1990). Motivational and Self-regulated Learning Components of Classroom Academic Performance. In *Journal of Educational Psychology*, 82(1), 33-40.

Preziosi, M.A., Coane, J.H. (2017). Remembering that Big Things Sound Big: Sound Symbolism and Associative Memory. In *Cognitive Research: Principles and Implications*, 2, 1-21. https://doi.org/10.1186/s41235-016-0047-y

R Core Team (2023). R: A Language and Environment for Statistical Computing. [Computer program] Version 4.3.0. https://www.r-project.org/

Raftery, A.E. (1995). Bayesian Model Selection in Social Research. In *Sociological Methodology*, 25, 111-163.

Ratcliff, R., Smith, P.L., Brown, S.D. & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. In *Trends in Cognitive Science*, 20(4), 260-281.

Rojczyk, A. (2011). Sound Symbolism in Vowels: Vowel Quality, Duration and Pitch in Sound-to-Size Correspondence. In *Poznań Studies in Contemporary Linguistics*, 47(3), 602-615.

Samarin, W.I. (1978). Linguistic Adaptation to Speech Function. In McCormack, W., Wurm, S.A. (a cura di), *Approaches to Language: Anthropological Issues*. The Hague: Mouton, 559-614.

Samuel, S., Cole, G. & Eacott, M.J. (2019). Grammatical Gender and Linguistic Relativity: A Systematic Review. In *Psychonomic Bulletin & Review*, 26(6), 1767-1786.

Sanchez, K., Hay, J. & Nilson, E. (2015). Contextual Activation of Australia Can Affect New Zealanders' Vowel Productions. In *Journal of Phonetics*, 48, 76-95.

Sapir, E. (1929). A Study in Phonetic Symbolism. In *Journal of Experimental Psychology*, 12(3), 225-239.

Schertz, J. (2013). Exaggeration of Featural Contrasts in Clarifications of Misheard Speech in English. In *Journal of Phonetics*, 41(3-4), 249-263.

Scobbie, J.M. (2006). Flexibility in the Face of Incompatible English VOT Systems. In Goldstein, L., Whalen, D.H. & Best, C.T. (a cura di), *Papers in Laboratory Phonology 8: Varieties of Phonological Competence*. Berlin: Mouton de Gruyter, 367-392.

Shinohara, K., Uno, R. (2022). Exploring the Positional Effects in Sound Symbolism. The Case of Hardness Judgments by English and Japanese Speakers. In *Languages*, 7(3), 1-11. https://doi.org/10.3390/languages7030179

Sidhu, D.M., Pexman, P.M. (2015). What's in a Name? Sound Symbolism and Gender in First Names. In *PLoS ONE*, 10(5), 1-22. https://doi.org/10.1371/journal.pone.0126809

Sidhu, D.M., Pexman, P.M. (2018). Five Mechanisms of Sound Symbolic Association. In *Psychonomic Bulletin & Review*, 25(5), 1619-1643.

Sidhu, D.M., Pexman, P.M. (2019). The Sound Symbolism of Names. In *Current Directions in Psychological Science*, 28(4) 398-402.

Silverstein, M. (1994). Relative Motivation in Denotational and Indexical Sound Symbolism of Wasco-Wishram Chinookan. In Hinton, L., Nichols, J. & Ohala, J.J. (a cura di), *Sound Symbolism*. Cambridge: University Press, 40-60.

Sullivan, L. (2020.) Universal and Cross-linguistic Gender-based Sound Symbolism in Korean Given Names. Generals Paper 2. University of Toronto, Toronto.

Svantesson, J-O. (2017). Sound Symbolism: The Role of Word Sound in Meaning. In *WIREs Cognitive Science*, 8(5), 1-12. https://doi.org/10.1002/wcs.1441

Tordini, O., Galatà, V., Avesani, C. & Vayra, M. (2018). Sound Maintenance and Change: Exploring Inter-language Phonetic Influence in First-generation Italo-Australian Immigrants. In Vietti, A., Spreafico, L., Mereu, D. & Galatà, V. (a cura di), *Speech in the Natural Context. Studi AISV 4*. Milan: Officinaventuno, 77-98.

Trowbridge J.E., Mintzes, J.J. (1988). Alternative Conceptions in Animal Classification: A Cross-Age Study. In *Journal of Research in Science Teaching*, 25(7), 547-571.

Usnadze, D. (1924). Ein experimenteller Beitrag zum Problem der psychologischen

Vigliocco, G., Vinson, D.P., Paganelli, F. & Dworzynski, K. (2005). Grammatical Gender Effects on Cognition: Implications for Language Learning and Language Use. In *Journal of Experimental Psychology: General*, 134(4), 501-520.

Wang, X. (2022). Recherches sur le symbolisme phonétique des noms propres. In *Signifiances (Signifying)*, 6(1), 98-120.

Winter, B., Oh, G.E., Hübscher, I., Idemaru, K., Brown, L., Prieto, P. & Grawunder, S. (2021). Rethinking the Frequency Code: A Meta-Analytic Review of the Role of Acoustic Body Size in Communicative Phenomena. In *Philosophical Transactions B*, 376, 1-13. https://doi.org/10.1098/rstb.2020.0400

Woschitz, J. (2019). Language in and out of Society: Converging Critiques of the Labovian Paradigm. In *Language & Communication*, 64, 53-67.

Zimman, L. (2023). Language, Gender, Race, and Sexuality: Intersectional Perspectives. In Duranti, A., George, R. & Riner, R.C. (a cura di), *A New Companion to Linguistic Anthropology*. Hoboken/Chichester: Wiley Blackwell, 525-541.

DOI: 10.17469/O2111AISV000016

ANNA ANASTASENI, ANTONIO ROMANO, SYLVAIN GERBER

Approssimanti e laterali palatali in italiano: l'effetto della pronuncia sulla conversione fono-grafemica in un gruppo di giovani parlanti torinesi

Approximants and palatal laterals in Italian: the effect of pronunciation on phono-graphemic conversion in a group of young speakers of Turin

Due to the articulatory proximity, Italian /ʎ/ and /j/ are confused by some speakers. The neutralization of the /ʎ/-/j/ contrast lets prevail [j]-like sounds. In order to investigate these phenomena, two classification tasks were proposed aiming to assess the incidence of the duration factor in distinguishing intervocalic segments. Furthermore, a reading task and a pseudo-words dictation task were proposed. It emerged that the duration plays a significant role in the identification of /ʎ/ and /j/. Through the analysis of participants' oral productions, two groups were identified: those who neutralize/ʎ/ and /j/ in pronunciation (NG: 30 %) and those who are able to distinguish them (DG). It emerged that the NG has greater difficulties in the grapho-phonemic conversion of pseudo-words than the GD. For the NG, it appears that different parameters (the quality of the sound and its duration) do not conflict in the recognition of [j] with non-prototypical duration, but that the speakers rely only on the duration of the sound.

Keywords: Auditory phonetics, identification task, palatal lateral, palatal approximant.

1. *Introduzione*

L'inventario fonologico dell'italiano distingue la laterale palatale /ʎ(ʎ)/(fonema intrinsecamente geminato in posizione postvocalica, che rappresenteremo da qui in avanti semplicemente come /ʎ/) e l'approssimante palatale /j/ (le cui realizzazioni nella stessa posizione possono presentare una durata meno vincolata)[1]; tuttavia tale distinzione è a basso rendimento funzionale.

[1] Riguardo alla scelta tra /ʎ/ o /ʎʎ/ (o /ʎː/) come fonema sottostante alle realizzazioni scempia (in contesto /#_, /C_) e geminata (in contesto V_V) della laterale palatale si veda Romano & Miletto (2017) e Marotta & Nocchi (2001). La durata di /ʎ/ supera in italiano standard e in diverse varietà regionali i 100 ms (Celata 2006); tuttavia non raggiunge la lunghezza della corrispondente geminata lessicale non palatale [lː]. Inoltre nei dati di Endo & Bertinetto (1999), [ʎː] sembrerebbe non indurre la vocale precedente ad accorciarsi come negli altri casi di sillaba chiusa. Riguardo a /j/ ci riferiamo invece a Romano & Miletto (2017: 66) che sottolineano come /j/ sia intrinsecamente breve "anche se [...] si presenta soggetto a variazione persino nella pronuncia dei professionisti della dizione". Precisiamo che nello studio di Anastaseni (2022) si considera un fonema /j/ dell'italiano, pur nella consapevolezza "del particolare statuto fonologico delle approssimanti [j] e [w]" (per approfondire la questione v. Marotta & Nocchi 2001: 52-53). Per queste, infatti, occorre mettere in conto varianti con

La laterale ha un complesso meccanismo articolatorio (Bladon & Carbonaro 1978; Canepari 1999; Quilis & Fernández 1969; Recasens 1984), e compare tardi nell'inventario fonologico dei bambini italofoni, per questo viene spesso semplificata (Galatà, V., Meneguzzi, G., Conter, L., Zmarich, C., 2012; Zampaulo & Haug, 2015). Inoltre, il fonema /ʎ/ è raro nelle lingue del mondo; è presente solo in 20 delle 451 considerate in UPSID (*UCLA Phonological Segment Inventory Database*) e ricorre solo in 16 lingue nel cui inventario compaia anche l'approssimante palatale /j/ (presente, al contrario in ben 378 lingue del database). Una perdita di contrasto tra /ʎ/ e /j/ è infatti favorita dalla somiglianza acustica, come mostra l'evoluzione dei sistemi fonologici di altre aree romanze in cui è endemico (Porras 2013)[2]. Diversamente dallo spazio accordato alle proprietà articolatorie di questi suoni (Recasens, D., Farnetani, E., Fontdevila, J., & Pallarès, M.D. 1993, Calamai & Bertinetto 2006), nella maggior parte della manualistica sperimentale è data una descrizione acustica solo di [ʎ], spesso senza esplicitare i parametri che lo distinguerebbero da [j] (in molti casi non trattato per via del suo incerto statuto fonologico). Data la debole possibilità di una distinta caratterizzazione acustica tra /ʎ/ e /j/ e la generale tendenza alla neutralizzazione del contrasto tra i due fonemi in certe aree (storicamente alcuni focolai nel sud Italia; cfr. Canepari, 1999), è stato avviato un progetto di studio acustico e percettivo. L'obiettivo principale è indagare i confini categoriali dell'approssimante palatale e della laterale palatale e le caratteristiche acustiche (sia in produzione che in percezione) che aiutano a distinguerle.

Si è scelto di focalizzare la nostra attenzione su parlanti torinesi d'italiano, per verificare se, in una comunità nella quale tradizionalmente non è documentata una neutralizzazione dell'opposizione tra questi suoni, si attesti tra i parlanti un certo grado di confusione e se questo abbia delle conseguenze nella conversione fono-grafemica.

Lo studio ha previsto quattro diverse fasi: (1) Somministrazione di due *task* di identificazione percettiva (2) Lettura di un breve brano (3) Dettato con ripetizione di pseudo-parole (4) *Task* di discriminazione percettiva (di tipo AB).

In questa sede soffermeremo l'attenzione sul *task* di dettato e in specifico sul rapporto tra pronuncia e accuratezza nella conversione fono-grafemica.

2. *Task di identificazione percettiva: uno studio preliminare*

Come menzionato, per valutare la capacità di categorizzare i due foni sono stati proposti preliminarmente due *task* di identificazione percettiva a un campione di 113 studenti che frequentano l'Università di Torino. In base ai risultati ottenuti sono stati strutturati i *task* successivi di cui qui vengono riassunti solo gli aspetti chiave (per maggiori dettagli sui due *task* di identificazione vd. Anastaseni, 2023).

realizzazione maggiormente vocalica (con fasi di stabilità timbrica e passaggio sufficientemente largo da consentire un buon irraggiamento della sonorità vocale).

[2] La progressione dei fenomeni che portano alla perdita di /ʎ/ in una lingua è considerata nel quadro dei mutamenti fonetici descritti da Fónagy (1983: 231).

L'obiettivo principale dei *task* era di valutare l'incidenza del fattore durata nella distinzione di segmenti intervocalici di tipo [ʎ], [j], [lj] e [lːj].

Si è scelto utilizzare una lista di pseudo-parole affinché le variabili lessicali non incidessero nel riconoscimento dei foni (vd Tabella 1)[3]. Partendo dai logatomi originari, pronunciati da due speaker professionisti, sono stati costruiti stimoli con segmenti intervocalici di durata diversa avendo cura di preservare le porzioni di transizione da un fono all'altro[4]. Gli stimoli sono stati manipolati facendo variare la durata dei foni target [ʎ] e [j] in un continuum di: 260 ms, 240 ms, 180 ms, 140 ms, 100 ms.

Tabella 1 - *Stimoli selezionati per il task di identificazione percettiva*

	ʎ	*j*	*lj*	*lːj*
[ˈa]_[a]	ciaglia	ciaia	cialia	ciallia
[ˈa]_[o]	naglio	naio	nalio	nallio
[ˈɔ]_[o]	roglio	roio	rolio	rollio
[ˈɔ]_[a]	toglia	toia	tolia	tollia

Il *task* consisteva nell'associare la pseudo-parola ascoltata con una etichetta ortografica in una lista chiusa di proposte ed è stato somministrato tramite la piattaforma Folerpa (Fernández Rei, E., Cajiao, A.A., Peláez, C.O., Garabal, J.A.C., 2021).

I partecipanti hanno avuto più difficoltà nella classificazione delle realizzazioni di /ʎ/ e /j/, rispetto a quelle di /lj/ e /llj/. Essi/e infatti sono stati/e capaci di associare correttamente le rese di /lj/ e /llj/ alle corrispondenti etichette ortografiche ‹li› e ‹lli›. Il grado di associazione /lj/~‹li› (92 % risposte target) e /llj/~‹lli› (91 %) è più alto (χ^2= 606,696; V=0,871) rispetto a /ʎ/~‹gl(i)› e /j/~‹i› (χ^2= 191,701; V= 0,490).

Inoltre, poiché è stato modultato il valore di durata dei foni [ʎ] e [j], è stato possibile osservare come tale parametro incida sul riconoscimento dei foni.

Ne è risultato che la durata ha un ruolo significativo nel corretto riconoscimento di /ʎ/ (*r* di Pearson = 0,947**, p = 0,01): al di sotto dei 140-160 ms il fono percepito è infatti maggioritariamente ricondotto a una risposta di tipo ‹i› (associata a /j/). C'è

[3] Nella scelta delle pseudo-parole sono state evitate parole esistenti, toponimi o nomi propri di persone o personaggi conosciuti. Un criterio è stato anche la non somiglianza con parole già esistenti, non è stato tuttavia possibile calcolare il numero di vicini ortografici, perché l'algoritmo messo in passato a disposizione dell'università di Padova per il calcolo dei vicini ortografici anche di parole inesistenti, era ed è al momento inattivo. Inoltre si è scelto di evitare la presenza di vocali palatali, che avrebbero potuto avere un effetto sulla classificazione dello stimolo. In questa sede non è stato possibile operare un'analisi statistica differenziata per contesto vocalico, si rimanda ad un lavoro futuro.

[4] È stata svolta una analisi acustica delle non parole registrate in cabina silente con due professionisti della dizione (un uomo, MG, e una donna, LD). Le realizzazioni delle due consonanti risultano molto oscillanti, da un punto di vista spettrale è difficile definire dei parametri fissi per entrambi gli speacker. La distinzione tra le due articolazioni i questi dati sembra dipendere maggiormente da un contrasto di durata. Per maggiori dettagli sull'analisi vedi Anastaseni A., Contini M., De Iacovo V. (2024).

un rapporto direttamente proporzionale tra la durata di [ʎ] e l'associazione target: maggiore è la durata del fono migliore è il suo riconoscimento. Invece, un rapporto inversamente proporzionale appare nel riconoscimento di [j] (*r* di Pearson = -0,988 p = 0,002): minore è la durata del fono maggiore è la sua percentuale di riconoscimento.

Per quanto riguarda le durate dei progredienti di un fono originariamente di tipo [j] sembrerebbe esserci una soglia intorno ai 200 ms: il fono percepito se la durata è superiore a questo valore non è però sistematicamente riconosciuto come quello corrispondente alla grafia ‹gl›. Di fronte a stimoli con [j] di durata imprevista (maggiore di 200 ms) la categorizzazione dello stimolo è incerta, come dimostrano le percentuali di risposte target intorno al 50 %.

Questo dato suggerisce che entrino in conflitto parametri diversi: la qualità del fono fa propendere i partecipanti verso l'associazione con la stringa ortografica ‹i›; d'altro canto però la durata così estesa non è prototipica di un fonema intrinsecamente breve (Romano & Miletto, 2017: 66) e spinge ad associare [j:] a ‹gl›.

Visti gli esiti di questi due primi *task* pilota, si è scelto di approfondire ulteriormente il tema.

3. *Task di dettato con ripetizione di pseudo-parole*

Poiché dal *task* di identificazione percettiva è emerso che il tratto di durata influenza la classificazione dei due fonemi /ʎ/ e /j/ e che esistono delle zone di ambiguità nel loro riconoscimento, si è deciso di proporre un *task* di dettato. Il vantaggio di questo compito è di superare il limite del *task* di identificazione di forzare il partecipante alla scelta di una stringa ortografica in una lista chiusa di proposte; infatti, la stessa struttura del *task* circoscrive le possibilità del locutore e può influenzare i risultati. Invece, il dettato consente lasciare scegliere la soluzione grafica preferita senza limitazioni. Anche il compito di dettato, tuttavia, impone di classificare i foni ascoltati per poterli trascrivere.

Inoltre, si è proposto un dettato al fine di valutare in che modo la classificazione influenza un compito di scrittura. Nel caso di ambiguità nella identificazione di due fonemi come distinti, il sistema di conversione fono-grafemicoo è più opaco; il che comporta che a un suono possono potenzialmente corrispondere numerose trascrizioni diverse. Questo ha importanti conseguenze sul modo di processare l'informazione ortografica (Angelelli, P., Marinelli, C.V., Putzolu, A., Notarnicola, A., Iaia, M., Burani, C, 2018). È interessante domandarsi se uno stimolo che presenta [j] o [ʎ] si debba considerare come ortograficamente non trasparente, in quanto la conversione fono-grafemica non è del tutto biunivoca.

Per queste ragioni si è scelto di proporre un *task* di dettato di pseudo-parole con manipolazione della durata del fono bersaglio ([j] e [ʎ]). Sono state selezionate pseudo-parole in modo che l'unica strategia disponibile di processamento ortografico sia la conversione fono-grafemica (la via indiretta).

Il *task* è strutturato in due fasi: dopo l'ascolto di ciascuna pseudo-parola il partecipante era tenuto prima a ripeterla ad alta voce, poi a scriverla su tastiera.

Si è ritenuto importante, infatti, il confronto tra produzione orale e produzione scritta. Pur consapevoli che la produzione orale non sia uno specchio preciso dell'organizzazione dell'inventario fonologico, si è deciso di indagare la correlazione tra capacità di distinguere a livello articolatorio e il comportamento grafico.

La nostra ipotesi è che l'impossibilità di distinguere i due foni in pronuncia possa corrispondere a una maggiore difficoltà nella classificazione dei fonemi (la cui distinzione per altro è a basso rendimento funzionale) e che questo possa portare a una più incerta conversione fono-grafemica. La predizione è quindi che i due fattori correlino tra loro: coloro che neutralizzano nella pronuncia i due foni (GN) avranno percentuali di accuratezza ortografica più basse che coloro in grado di distinguerli (GD):

(1) GD>GN

Per questo sono stati raccolti dati non solo di produzione scritta ma anche orale. (§ 3.3.1.).

3.1 Metodologia

È stato proposto un *task* di dettato in cui è stato richiesto ai partecipanti di ripetere ogni stimolo prima di scriverlo su tastiera e un *task* di lettura di un breve brano. Il brano contiene 16 parole che presentano i foni bersaglio ([j] e [ʎ:]) in contesto intervocalico in posizione pre- e post-accentuale.

Tutti i dati sono stati raccolti presso il Laboratorio di Fonetica Sperimentale "Arturo Genre" dell'Università degli Studi di Torino e le registrazioni eseguite in una cabina silente, con l'ausilio di un microfono dinamico Shure SM58 e la scheda audio Focusrite Scarlett 4i4.

3.1.1 Partecipanti

I due *task* sono stati proposti a 60 parlanti torinesi dai 15 ai 30 anni, divisi in classi di età (15-20](20-25](25-30], venti partecipanti per ciascuna.

Sono state selezionate solo persone prevalentemente monolingui e che hanno indicato come regione di provenienza della propria famiglia il Piemonte. È stato chiesto loro di indicare tutte le lingue di cui hanno competenza attiva e passiva e il rispettivo livello. Tutti i partecipanti al *task* hanno dichiarato di non avere problemi di udito, motori, di vista non corretti e di aver mai sospettato una ipoacusia.

Si è scelto di proporre il *task* a giovani adulti (dai 15 ai 30 anni) ipotizzando che la neutralizzazione dell'opposizione tra /ʎ/ e /j/ sia un fenomeno in espansione, in particolare in Piemonte dove non ci si aspetta questo tratto come autoctono.

Ai partecipanti sono stati proposti un questionario sociolinguistico e un questionario di autovalutazione volto a identificare difficoltà di letto-scrittura: il Questionario Vinegrad Plus secondo la versione modificata di Montesano & Valenti (2017)[5].

[5] Si tratta di un adattamento in italiano della *Revised Adult Dyslexia Checklist* (ADCL) di Vinegrad (1994), un questionario *self-report* di 26 domande volto all'autovalutazione delle proprie difficoltà di letto-scrittura e calcolo. Non è uno strumento diagnostico, ma permette di identificare chi ha difficoltà riconducibili a un Disturbo Specifico dell'Apprendimento. Si considera a rischio, infatti, chi

3.1.2 Materiale

Sono stati dettati quattro gruppi di pseudo-parole con foni bersaglio [ʎː], [j] in posizione intervocalica: [ˈa]_[a]; [ˈa]_[o]; [ˈɔ]_[o]; [ˈɔ]_[a].

Tutti gli stimoli presentano la stessa struttura fonotattica: sono pseudo-parole bisillabiche e parossitone; mantenere costante queste caratteristiche è necessario per evitare eventuali effetti di lunghezza dello stimolo (verificati in scrittura per l'italiano da Marinelli, C.V., Romani, C., Burani, C., & Zoccolotti, P. (2015) e di posizione dell'accento. Inoltre, è stata evitata la presenza di vocali palatali poiché uno studio condotto sullo spagnolo proprio riguardo alla distinzione tra /ʎ/ e /j/ ha mostrato che potrebbero condizionare l'identificazione in favore di [j] (Rost Bagudanch 2014).

Tabella 2 - *Stimoli selezionati per il task di dettato*

	ʎ	*j*
[ˈa]_[a]	ciaglia	ciaia
[ˈa]_[o]	naglio	naio
[ˈɔ]_[o]	roglio	roio
[ˈɔ]_[a]	toglia	toia

Si è scelto di utilizzare parole inesistenti per evitare che variabili lessicali incidessero sulla prova. Inoltre, in questo modo l'unica strategia possibile da adottare per scrivere è l'utilizzo della via indiretta, di conversione fono-grafemica.

Gli stimoli scelti sono riportati nella Tabella 2.

3.1.2. Trattamento degli stimoli

Una volta selezionate le pseudo-parole per il *task*, sono stati registrati due speaker professionisti in una cabina silente, che hanno ripetuto gli stimoli in isolamento.

In seguito, sono state uniformate con l'ausilio del software di editing audio GoldWave le durate delle vocali antecedenti ai foni target [ʎː] e [j] a 145 ms (± 5 ms) e 160 ms (± 5 ms).

Si è poi calcolato il valore medio della durata di [ʎː] nelle quattro pseudo-parole (*ciaglia*, *naglio*, *roglio*, *toglia*): 158 ms. Sono quindi stati uniformati a questo valore tutti gli stimoli (con un'approssimazione di 2 ms). In seguito, mantenendo la vocale precedente identica si è fatta variare la durata del fono target; in questo modo si è creato un *continuum* di durata di [ʎ] in passi da 40 ms ([ʎ] di durata 260 ms, 180 ms e 100 ms).

È stata fatta la stessa operazione con gli stimoli che presentano il fono [j], nell'audio originale di durata media di 180 ms, manipolato in modo da riprodurre lo stesso *continuum* di [ʎ] (260 ms, 180 ms, 100 ms).

totalizza un punteggio superiore o uguale a 9 punti (Cornoldi & Montesano, 2020; Montesano & Valenti, 2017; Montesano & Valenti, 2018).

In questo modo si è potuto sottoporre a ciascun partecipante l'ascolto di stimoli esattamente identici fatta eccezione per la durata del fono target. Questo permette di fare alcune interessanti osservazioni sulla sensibilità dei partecipanti al test rispetto a questo tratto di durata.

3.2. Procedura sperimentale

Il *task* consiste in un dettato di pseudo-parole. Ciascuno stimolo (24 in totale, vedi Tabella 2), dopo essere stato ascoltato in cuffia, veniva ripetuto dal soggetto e poi scritto. In questo modo è stato possibile il confronto tra produzione orale e resa grafica.

I soggetti hanno scritto al computer per limitare gli errori legati a difficoltà di tipo grafo-motorio. Sono stati tenuti in conto in fase di analisi eventuali errori di battitura e le specificità della scrittura su tastiera (Dhakal, V., Feit, A.M., Kristensson, P.O., & Oulasvirta, A., 2018).

Per la somministrazione del *task* si è scelto di utilizzare il software OpenSesame, che permette di raccogliere anche i tempi di risposta.

Prima della somministrazione del *task* è stato fatto leggere ai partecipanti un breve testo in cui sono state inserite sedici parole che presentano i foni bersaglio [ʎ] e [j] sia in posizione intervocalica pre-accentuale e post-accentuale. In questo modo, è stato possibile valutare la pronuncia dei partecipanti, in particolare la capacità di distinguere la laterale palatale e l'approssimante palatale.

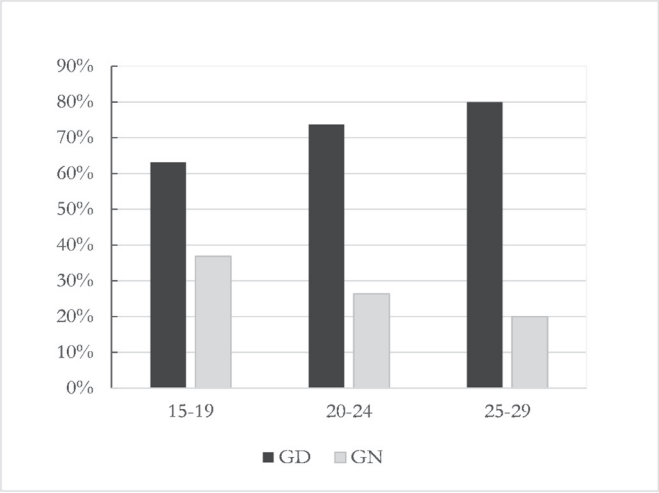
3.3. Analisi dei dati e discussione

3.3.1. Produzione orale

In primo luogo sono state analizzate con il software Praat le registrazioni della lettura ad alta voce e delle ripetizioni di pseudo-parole nel *task* di dettato. A partire da questi dati, i partecipanti sono stati divisi in due gruppi: coloro in grado di distinguere in produzione i foni [ʎ] e [j] (GD) e coloro che li confondono estendendo la pronuncia [j] (GN).

Ne è emerso che in tutte le classi di età la percentuale di partecipanti che neutralizza (GN) è superiore al 25 % (Figura 1). Se ne deduce che questo fenomeno è diffuso anche in Piemonte, dato rilevante perché inatteso, non essendo un tratto che tradizionalmente viene attribuito come caratteristica propria dell'italiano di questa regione. Inoltre si nota che la percentuale di persone che neutralizzano l'opposizione è più alta tra i più giovani; la neutralizzazione in pronuncia sembra quindi in leggera espansione. Sarebbe interessante ampliare il campione aggiungendo classi di età diverse, in modo da poter valutare con più precisione come si distribuisce questa caratteristica nella popolazione.

Figura 1- *Pronuncia dei soggetti per classe di età: Percentuale di soggetti che presentano una pronuncia che distingue (GD) o neutralizza l'opposizione /ʎ/ ~ /j/(GN) divisi per classe di età*



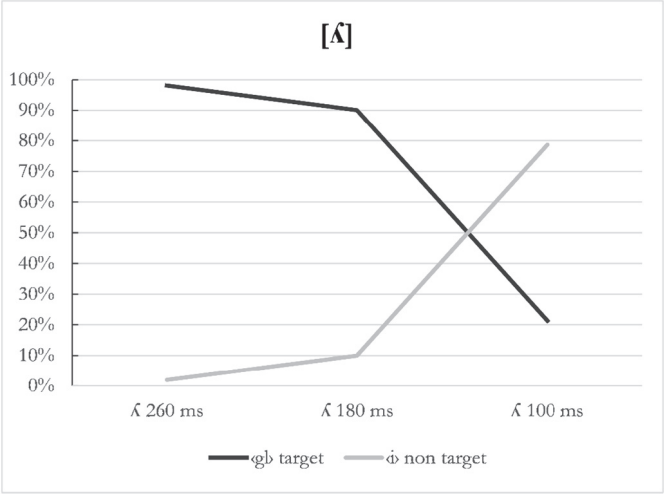
3.3.2. Produzione scritta

Per quanto riguarda invece l'accuratezza ortografica nel *task* di dettato, i risultati sono fortemente in linea con quelli del *task* di identificazione percettiva.

Per semplificare lo studio quantitativo dei risultati, si è scelto in questa sede di considerare la variabile *risposta* sempre come binaria: *target* o *non target*; sono state considerate *target* rispettivamente le corrispondenze [ʎ] ~ ‹gl(i)› e [j] ~ ‹i›. Si noti però che poiché si tratta di un *task* di dettato libero (in cui non sono state date indicazioni sulle grafie da utilizzare) sono state adottate da parte di alcuni partecipanti anche soluzioni grafiche creative[6]. Si è scelto qui di assimilare a ‹i› le grafie ‹j›, ‹ij›, ‹ji›, ‹y›, ‹ii› (14 occorrenze/373 totali).

[6] Le soluzioni grafiche adottate come *ciaiia, ciajia, ciaija, ciaya* nella trascrizione sono state adottate solo da partecipanti del gruppo GD maggiormente in risposta agli stimoli con [j] maggiore di 180 ms. Questo può indicare che i partecipanti hanno rilevato un'anomalia negli stimoli proposti. Se ne deduce quindi che tali partecipanti abbiano la capacità di distinguere l'approssimante in base a indici acustici, a prescindere dalla sua durata. Si rimanda a un'altra sede una discussione più dettagliata di tale fenomeno per mancanza di spazio qui.

Figura 2 - *Percentuali di realizzazione target e non target nella scrittura sotto dettatura di pseudo-parole che presentano [ʎ] di durata 260 ms, 180 ms e 100 ms*



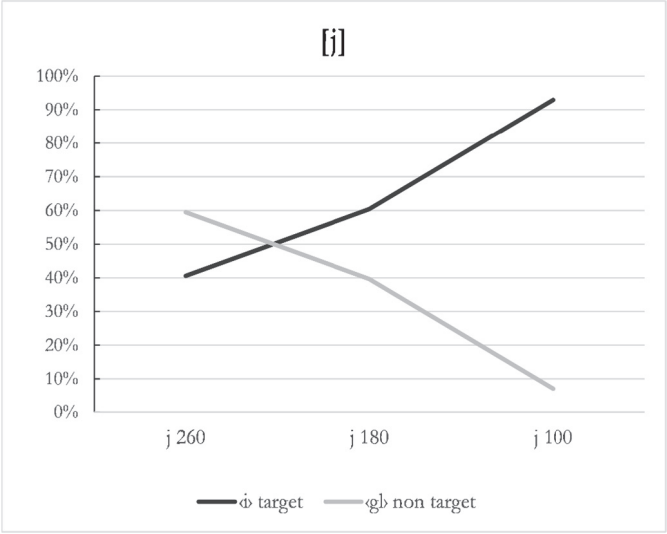
Parallelamente a quanto già notato per il *task* di identificazione percettiva, le percentuali di riconoscimento di [ʎ] (quindi di corretta conversione fono-grafemica) si abbassano al diminuire della durata del fono, fino ad arrivare al 21 % quando ha durata di 100 ms (Figura 2). Sembrerebbe che al di sotto della soglia virtuale di 140 ms (verificata nel *task* di identificazione percettiva, § 2) avvenga una ricategorizzazione fonologica.

Sebbene con un margine di incertezza, [ʎ] a 100 ms viene associato a grafie di tipo ‹i› (Figura 2).

Viceversa, per quanto riguarda [j], il profilo è inverso (Figura 3): maggiore è la durata del fono, minore è il suo riconoscimento. Come già notato per il *task* di identificazione percettiva (§ 2), anche qui è importante osservare che per [j] con durata di 260 ms i valori minimi di accuratezza raggiunti si aggirano intorno al 50 %, a differenza di [ʎ] per cui i valori sono più bassi (21.33 %). Questo dato indica che per gli stimoli [j] con durata insolita i partecipanti mostrano un alto grado di incertezza nella categorizzazione.

Il *task* di dettato ha confermato i risultati dei *task* di identificazione (§ 2), mostrando che il tratto di durata gioca un ruolo fondamentale anche per la corretta conversione fono-grafemica di /ʎ/ e /j/. L'associazione [ʎ]-durata lunga e [j]-durata breve guida anche i processi di scrittura.

Figura 3 - *Percentuali di realizzazione target e non target nella scrittura sotto dettatura di pseudo-parole che presentano [j] di durata 260 ms, 180 ms e 100 ms*



In particolare, [ʎ] di durata 100 ms sembra non essere compatibile con la classificazione /ʎ/. Del resto, come dimostrato da Mairano & De Iacovo (2020), anche le varietà di italiano settentrionale conoscono la geminazione intrinseca[7]; per questo una durata così ridotta non rientra tra le caratteristiche attribuite a /ʎ/ e il fono viene ricategorizzato. D'altro canto, per quanto riguarda [j:], l'incertezza va ricondotta alla presenza di due parametri contrastanti che spingono da un lato all'associazione con grafie di tipo ‹i› (la qualità acustica del fono) dall'altro alla grafia ‹gl(i)› (la durata lunga).

Dopo questa prima analisi complessiva, i dati sull'accuratezza ortografica sono stati incrociati con quelli sulla pronuncia dei partecipanti.

Ne è emerso che, come atteso, coloro che neutralizzano i due fonemi nella realizzazione orale (GN) hanno anche maggiori difficoltà nella conversione fono-grafemica delle pseudo-parole (56,6 % risposte target) rispetto a coloro che sono in grado di distinguere (GD) i due foni in pronuncia (72,4 % risposte target).

Questo dato supporta l'idea che per il GN le corrispondenze [ʎ]~‹gl(i)› e [j]~‹i› non siano del tutto trasparenti e la loro classificazione come fonemi distinti sia più difficoltosa.

Analizzando nel dettaglio le percentuali di risposte target distribuite per fono bersaglio si osservano i risultati di maggiore interesse. Nel paragrafo che segue si riportano nei dettagli la metodologia di analisi utilizzata e i risultati più significativi.

[7] Il corpus analizzato comprende parlanti delle città di: Torino, Genova, Milano, Bergamo, Parma, Venezia.

3.3.1 L'influenza della pronuncia sull'accuratezza ortografica in stimoli con [ʎ] e [j] con durata diversa

Si è scelto di studiare l'impatto delle due variabili *gruppo* (nelle due modalità GN o GD) e *stimolo* (nelle sei modalità 100j, 180j, 260j, 100ʎ, 180ʎ, 260ʎ) e la loro interazione su una variabile dipendente a risposta binaria (*target* e *non target*).

Dato che ad ogni partecipante è stato chiesto di scrivere diversi stimoli nelle sei modalità (*ciaglia/ciaia*, *roglio/roio*, *toglia/toia* e *naglio/naio*), è stata introdotta la variabile *soggetto* come effetto casuale nel modello. Pertanto, considerata la natura della variabile dipendente (a risposta binaria e misure ripetute), abbiamo scelto di usare un modello di regressione logistica con effetti casuali. L'analisi è stata eseguita utilizzando la funzione *glmer( )* del pacchetto software *R lme4*. Si tratta di un test che confronta due modelli annidati. L'ipotesi nulla del test è che il modello con meno parametri sia sufficiente e che l'aggiunta di altri parametri non sia necessaria per spiegare meglio le variazioni della variabile risposta. La statistica del test corrisponde al doppio della differenza tra le *log-likelihood* dei due modelli e segue una distribuzione *Chi-quadro* il cui grado di libertà è pari alla differenza del numero di parametri (nel caso di effetti fissi). Se il valore *p* è inferiore alla soglia stabilita (0,05), viene rifiutata l'ipotesi nulla e si ritiene che il contributo dei parametri aggiuntivi sia significativo.

È stata esaminata la struttura degli effetti fissi. È stata poi eseguita una selezione *top-down* che consente di eliminare progressivamente i termini non significativi. In ogni fase, è stato eliminato dal modello l'effetto il cui valore di *p* associato al test del rapporto di verosimiglianza è il più grande e maggiore di 0,05. Il processo si interrompe quando tutti gli effetti sono stati rimossi dal modello o quando il valore di *p* è inferiore a 0,05.

I vari test della selezione *top-down* sono stati eseguiti utilizzando la funzione *anova( )* in R.

Per giudicare le prestazioni del modello finale, è stata utilizzata l'area sotto la curva ROC: *AUC* (area *under curve*), che assume valori compresi tra 0 e 1. A tale scopo, abbiamo utilizzato il pacchetto *AUC* del software R. Una spiegazione dettagliata della curva ROC si trova nel libro di Gilbert Saporta (2011).

Una volta ottenuto il modello finale, è stato applicato il metodo presentato nell'articolo di Hothorn, T., Bertz, F., Westfall, P. (2008) e la funzione *glht* del pacchetto *multcomp* del software R. Questo metodo garantisce che il rischio legato all'assunzione simultanea di tutte le decisioni non superi una soglia che abbiamo fissato in anticipo (0,05) aggiustando i valori di *p*.

Infine, è stato mantenuto nel modello l'effetto fisso dell'interazione tra i fattori *gruppo* e *stimolo* ($\chi^2(5)= 22{,}21$, $p<0{,}0001$), il che significa che l'impatto del fattore *gruppo* è diverso a seconda delle modalità del fattore *stimolo*. Inoltre, l'area *AUC* è pari a 0,91, un valore molto soddisfacente. Gli esiti di tutti i confronti effettuati sono riassunti in forma schematica nelle Appendici 1 e 2.

3.3.2 Risultati

In primo luogo, è stata confrontata la percentuale di risposte target per ogni stimolo sia nel GN che nel GD. È stato operato un confronto diretto entro le percentuali di accuratezza ortografica di ciascuno stimolo con i restanti aventi lo stesso fono target ma con durata differente (es. [ʎ] di durata 260 ms è stato confrontato sia con [ʎ] 180 ms che con [ʎ] 100 ms).

Ne è risultato che la differenza di accuratezza ortografica è statisticamente significativa per il confronto rispettivamente di: [ʎ] 100 ms e [ʎ] 180 ms, [ʎ] 100 ms e [ʎ] 260 ms, [j] 100 ms e [j] 180 ms, [j] 100 ms e [j] 260 ms; al contrario non è significativa nel confronto tra gli stimoli di durata 180 ms e 260 ms, fatta eccezione per gli stimoli con [j] nel GN (Figura 4 e 5; vd. anche Appendice 2).

In seguito è stata calcolata la differenza di accuratezza ortografica nei due gruppi (GD e GN) per stimolo (Appendice 2). Di qui i risultati più interessanti per le loro possibili interpretazioni.

Infatti, è emerso che essa non è statisticamente significativa per quanto riguarda gli stimoli nelle tre modalità di [ʎ] (GD≈GN). Il trattamento del fono è uniforme nei due gruppi (Figura 6, sinistra). Sia per il GD che per il GN il dato di durata prevale su quello di qualità: gli stimoli con durata di 180 ms e 260 ms sono saldamente associati al *target*, al contrario invece di quelli con durata di 100 ms.

Figura 4 - *Percentuali di realizzazione target nella scrittura sotto dettatura di pseudo-parole che presentano [ʎ] di durata 260 ms, 180 ms e 100 ms di GD e GN*

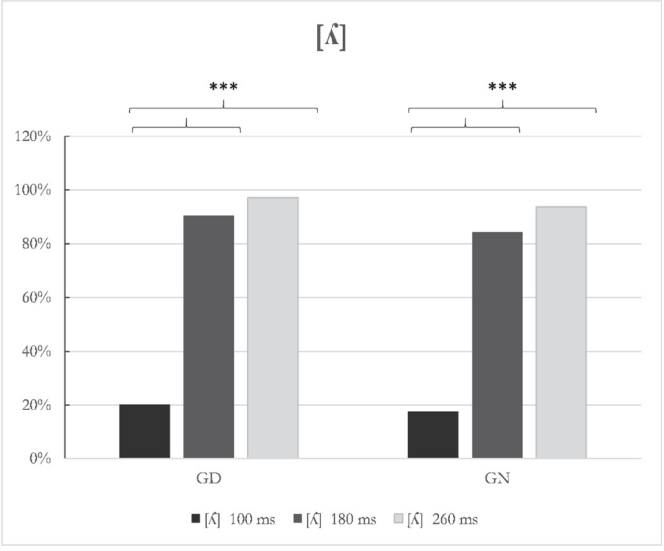
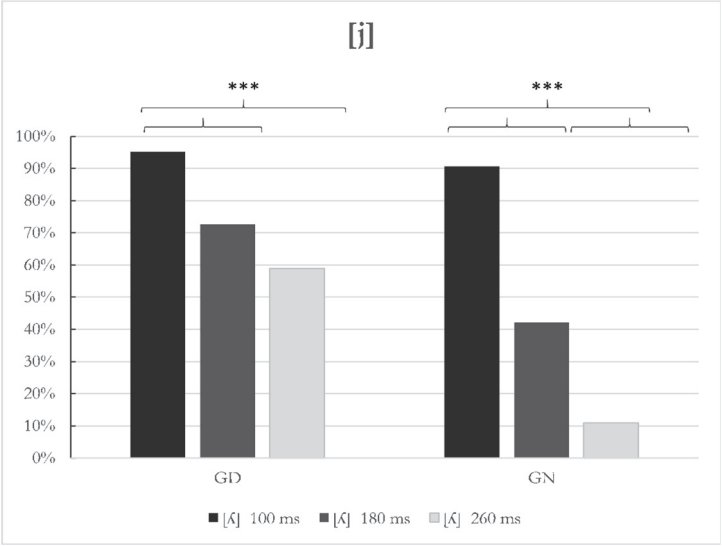
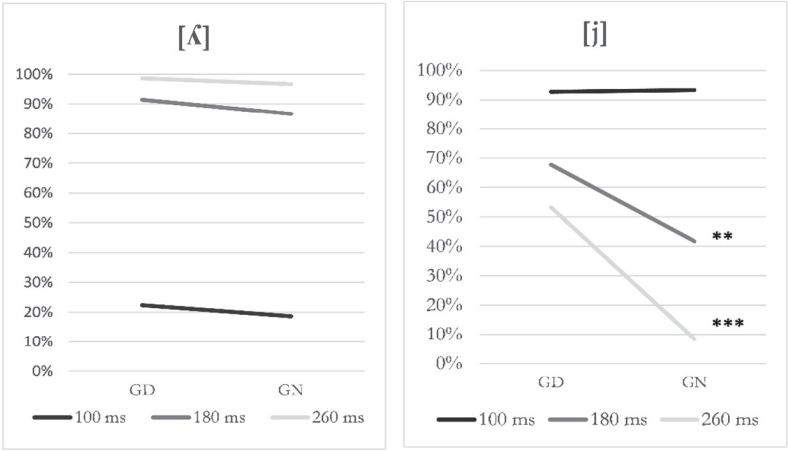


Figura 5 - *Percentuali di realizzazione target nella scrittura sotto dettatura di pseudo-parole che presentano [j] di durata 260 ms, 180 ms e 100 ms di GD e GN*



Viceversa, per quanto riguarda le pseudo-parole con [j] di durata di 180 ms e 260 ms la discrepanza tra i due gruppi è statisticamente significativa (180 j <0.01 **; 260 j <0.01 ***). La percentuale di risposte non target è sensibilmente più bassa presso il GN. Si noti, in aggiunta, che nel 94 % delle risposte non target la risposta data dai partecipanti è stata <gl(i)>. Per il gruppo di chi neutralizza sembra che non entrino in conflitto parametri diversi (la qualità del fono e la sua durata) nel riconoscimento di [j] con durata non prototipica, ma che i parlanti si affidino unicamente alla durata del fono per la sua classificazione, associando saldamente [j] 260 ms alla grafia <gl(i)>. A differenza di GD, per GN avviene una ricategorizzazione. Per quanto riguarda il GD invece le percentuali di accuratezza ortografica in [j] 260 ms si aggirano intorno al 50 %. I dati suggeriscono che i partecipanti che distinguono in pronuncia i due foni, non affidino la classificazione soltanto al parametro di durata del fono.

Figura 6 - Percentuali di realizzazione target e non target nella scrittura sotto dettatura di pseudo-parole che presentano rispettivamente [ʎ] (a sinistra) e [j] (a destra) di durata 260 ms, 180 ms e 100 ms



4. *Conclusioni e prospettive*

In conclusione, il *task* di dettato proposto ha permesso di riconfermare che il tratto di durata gioca un ruolo importante nella classificazione di /ʎ/ e /j/, come già verificato in un precedente *task* di identificazione. La medesima associazione [ʎ]-durata lunga e [j]-durata breve guida la conversione fono-grafemica.

Inoltre, l'analisi delle produzioni orali dei partecipanti ha dimostrato che la neutralizzazione in pronuncia dei due foni è diffusa tra i giovani parlanti piemontesi e sembra in espansione. Le ragioni di tale espansione si devono ricercare nelle dinamiche articolatorie complesse di [ʎ], nella debole opposizione fonologica dei due fonemi e nella vicinanza acustica tra i due foni.

Infine, la possibilità di confrontare i dati di produzione orale e scritta hanno avuto importanti conseguenze sia da un punto di vista metodologico che pratico.

È risultato infatti che le percentuali di accuratezza ortografica nella scrittura sotto dettatura di stimoli con [j] di durata 180 ms e 260 ms del gruppo di chi è in grado di distinguere i due foni in pronuncia (GD) e di chi li neutralizza (GN) siano significativamente diverse. Questo dato è interessante perché suggerisce che non vengano usati gli stessi parametri nella classificazione dei due fonemi. Il GN infatti sembra affidarsi unicamente al tratto di durata, al contrario del GD maggiormente condizionato anche dalla qualità del fono.

References

Anastaseni, A. & Romano, A. (2023). A perceptual experiment testing the Italian/ʎ/-/j/ contrast. In *20th International Congress of the Phonetic Sciences* (ICPhS) 236-240. Guarant International.

Angelelli, P., Marinelli, C.V., Putzolu, A., Notarnicola, A., Iaia, M., Burani, C. (2018). Learning to spell in a language with transparent orthography. Distributional properties of orthography and whole-word lexical processing. In *The Quarterly Journal of Experimental Psychology*, 71, 704-726, https://doi.org/10.1080/17470218.2016.1275715

Bertinetto, P.M., & Loporcaro, M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. In *Journal of the International Phonetic Association*, 35(2), 131-151, https://doi.org/10.1017/S0025100305002148

Bladon, R., & Carbonaro, E. (1978). Lateral consonants in Italian. In *Journal of Italian Linguistics*, 3(1), 43-54.

Calamai, S., & Bertinetto, P.M. (2006). Per uno studio articolatorio dei legamenti palatale, labiovelare e labio-palatale dell'italiano. In V. Giordani *et al.* (a cura di), *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche,* Torriana (RN): EDK, 43-56.

Canepari, L. (1999). *Il MaPI. Manuale di pronuncia italiana*. Bologna: Zanichelli [1ª ed. 1992].

Celata, C. (2006). *Analisi del processo di retroflessione consonantica in area romanza, con dati sperimentali dal còrso e dal siciliano.* Tesi di dottorato inedita, Scuola Normale Superiore.

Cornoldi, C., & Montesano, L. (2020). *Nuova batteria per studenti universitari e adulti LSC-SUA.* Trento: Erickson.

Dhakal, V., Feit, A.M., Kristensson, P.O., & Oulasvirta, A. (2018). Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 646, 1-12. https://doi.org/10.1145/3173574.3174220

Fernández Rei, E., Cajiao, A.A., Peláez, C.O., Garabal, J.A.C. (2021). FOLERPA: Ferramenta On-Line para ExpeRimentación PerceptivA, Santiago de Compostela: Instituto da Lingua Galega. https://ilg. usc. gal/folerpa (accessed on 20 October 2021).

Fónagy, I. (1983). *La vive voix. Essais de psycho-phonétique*. Paris: Payot (2e éd. 1991).

Galatà, V., Meneguzzi, G., Conter, L., Zmarich, C. (2012). Primi dati sull'acquisizione fonetico-fonologica dell'italiano L2 in prescolari rumeni. In Paoloni A. & Falcone M. (a cura di), *La voce nelle applicazioni*. Roma: Bulzoni, 35-50.

Hothorn, T., Bertz, F., Westfall, P. (2008). Simultaneous inference in general parametric models. In *Biometrical journal*, 50(3), 346-363.

Mairano, P., & De Iacovo, V. (2020). Gemination in Northern versus Central and Southern Varieties of Italian: A Corpus based Investigation. In *Language and Speech*, 63(3), 608-634.

Marinelli, C.V., Romani, C., Burani, C., & Zoccolotti, P. (2015). Spelling acquisition in English and Italian: A cross-linguistic study. Frontiers in psychology, 6, 1843.

Marotta, G., & Nocchi, N. (2001). La liquida laterale nel livornese. In *Rivista italiana di dialettologia*, 25, 285-326. doi: 10.1400/12765

Montesano, L., & Valenti, A. (2018). SLD: promoting continuity between School and University education. Results of a screening program in Calabria. In *Form@re - Open Journal per la formazione in rete*, 18(1), 192-206. http://dx.doi.org/10.13128/formare-22313

Pettorino, M., & Giannini, A. (1991). Indagine acustica sulle approssimanti dell'italiano. In *Atti del XIX convegno AIA,* Napoli, 10-12 aprile 1991, 441-447.

Porras, J.E. (2013). Spanish Yeísmo: A Cognitive Linguistic Approach to Phonological Change. In: Gómez R. & Molina Martos I. (a cura di), *Variación yeísta en el mundo hispánico*. Iberoamericana Editorial Vervuert, 335-352.

Quilis, A., & Fernández, J.A. (1969). *Curso de fonética y fonología españolas para estudiantes angloamericanos*. 4. Vol. 2. Madrid: Instituto "Miguel Cervantes".

Recasens, D., Farnetani, E., Fontdevila, J., & Pallarès, M.D. (1993). An electropalatograpic study of alveolar and palatal consonants in Catalan and Italian. In *Language and Speech*, 36(2-3), 213-234.

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. In *The Journal of the Acoustical Society of America*, 76(6), 1624-1635.

Romano, A., & Miletto, A.M. (2017). *Argomenti scelti di Glottologia e Linguistica*. Torino: Omega.

Romano, A., Anastaseni A., Contini M., De Iacovo V. (2024). Palatali approssimanti e laterali in Italia: un contributo preliminare di analisi acustica e percettiva, in P. Roseano & W. Elvira-García (a cura di), *Avances metodológicos en fonética y prosodia*, Madrid: Editorial UNED, 295-307.

Rost Bagudanch, A. (2014). El papel de la percepción en el yeísmo: Estudio preliminar con jueces catalanohablantes, *Loquens*, 1(2). doi: http://dx.doi.org/10.3989/loquens.2014.010

Rost Bagudanch, A. (2016). La percepción de /ʎ/ y /j/ en catalán y en español. Implicaciones en la explicación del yeísmo. In *Estudios de Fonética Experimental*, XXV, 40-80.

Saporta, G. (2011). *Probabilités analyse des données et Statistiques*, Editions TECHNIP, 622.

Zampaulo, A., & Haug, D. (2015). The evolution of the (alveolo) palatal lateral consonant in Spanish and Portuguese. In D. Haug (a cura di.), *Historical Linguistics 2013*, Amsterdam: John Benjamins, 69-86.

Software and statistical packages

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [Computer program] Version 6.1.40. https://www.praat.org

Bryer, J., Speerschneider, K. (2016). Likert: Analysis and Visualization Likert Items. [R package] Version 1.3.5. https://CRAN.R-project.org/package=likert

Falissard, B. (2012). Psy: Various procedures used in psychometry. [R package] Version 1.1. https://cran.r-project.org/package=psy

Appendice

1. Tabella riassuntiva relativa ai dati di confronto tra i due *gruppi* (GD e GN) per ciascuno *stimolo*
   Estimate: coefficienti di regressione.
   Std. Error: deviazioni standard.

Z: valore della statistica del test.
value Pr(>|z|): valore p del test.

	Estimate	Std. Error	z value	Pr(>|z|)
GD - GN | 100j == 0	0.8053	0.6026	1.336	0.9565
GD - GN | 100ʎ == 0	0.1411	0.4408	0.320	1.0000
GD - GN | 180j == 0	1.4203	0.3749	3.789	<0.01 **
GD - GN | 180ʎ == 0	0.6417	0.4850	1.323	0.9594
GD - GN | 260j == 0	2.6206	0.4825	5.431	<0.01 ***
GD - GN | 260ʎ == 0	0.8526	0.7214	1.182	0.9826

2. Tabella riassuntiva relativa ai dati di confronto tra ciascuna coppia di stimoli per il gruppo GD e per il gruppo GN.

100j - 100ʎ | GD == 0	4.7107	0.4296	10.964	<0.01 ***
100j - 180j | GD == 0	2.1231	0.4079	5.205	<0.01 ***
100j - 180ʎ | GD == 0	0.7663	0.4516	1.697	0.8072
100j - 260j | GD == 0	2.7990	0.4034	6.938	<0.01 ***
100j - 260ʎ | GD == 0	-0.4968	0.5827	-0.853	0.9990
100ʎ - 180j | GD == 0	-2.5876	0.2799	-9.245	<0.01 ***
100ʎ - 180ʎ | GD == 0	-3.9444	0.3487	-11.313	<0.01 ***
100ʎ - 260j | GD == 0	-1.9117	0.2657	-7.194	<0.01 ***
100ʎ - 260ʎ | GD == 0	-5.2075	0.5092	-10.226	<0.01 ***
180j - 180ʎ | GD == 0	-1.3568	0.3230	-4.200	<0.01 ***
180j - 260j | GD == 0	0.6759	0.2450	2.759	0.1353
180j - 260ʎ | GD == 0	-2.6199	0.4906	-5.340	<0.01 ***
180ʎ - 260j | GD == 0	2.0327	0.3170	6.412	<0.01 ***
180ʎ - 260ʎ | GD == 0	-1.2631	0.5274	-2.395	0.3118
260j - 260ʎ | GD == 0	-3.2958	0.4870	-6.767	<0.01 ***
100j - 100ʎ | GN == 0	4.0465	0.5585	7.245	<0.01 ***
100j - 180j | GN == 0	2.7380	0.5118	5.350	<0.01 ***
100j - 180ʎ | GN == 0	0.6026	0.5591	1.078	0.9918
100j - 260j | GN == 0	4.6142	0.6029	7.653	<0.01 ***
100j - 260ʎ | GN == 0	-0.4496	0.6787	-0.662	0.9999
100ʎ - 180j | GN == 0	-1.3085	0.4297	-3.045	0.0604
100ʎ - 180ʎ | GN == 0	-3.4438	0.4950	-6.957	<0.01 ***
100ʎ - 260j | GN == 0	0.5678	0.5291	1.073	0.9920
100ʎ - 260ʎ | GN == 0	-4.4961	0.6289	-7.149	<0.01 ***
180j - 180ʎ | GN == 0	-2.1353	0.4421	-4.830	<0.01 ***
180j - 260j | GN == 0	1.8763	0.4851	3.868	<0.01 **
180j - 260ʎ | GN == 0	-3.1876	0.5877	-5.424	<0.01 ***
180ʎ - 260j | GN == 0	4.0116	0.5445	7.367	<0.01 ***
180ʎ - 260ʎ | GN == 0	-1.0522	0.6289	-1.673	0.8219
260j - 260ʎ | GN == 0	-5.0638	0.6687	-7.573	<0.01 ***

DOI: 10.17469/O2111AISV000017

ROBERTA BIANCA LUZIETTI

La degeminazione delle rotiche in Toscana: indagine preliminare in un archivio orale storico

Degemination in Tuscan rhotics: a preliminary analysis in an historical oral archive

The paper presents the first results of a preliminary sociophonetic investigation on the residual phenomenon of rhotic degemination in the Tuscan Pratese variety, through the reuse of the Angela Spinelli historical oral archive, initially conceived for other (non-linguistic) purposes. In past research, some authors reported the presence of rhotic degemination phenomena in the nearby territories investigated by the historian. However, until recent times the phenomenon has been under-investigated. To find evidence of rhotic degemination in a phonetic laboratory would now be impossible since this phenomenon embodies the residual vestige of a past tendency, challenged by the diffusion of more standard-like pronunciations in favor of the maintenance of the singleton–geminate contrast. The assumption is that the consultation of the Spinelli archive would allow to find and study the use of past non-standard forms in situations where participants narrate of life-threatening events and their speech is, therefore, less controlled, and more spontaneous.

Keywords: Sociophonetics, oral history archives, Tuscan varieties, rhotic degemination.

1. *Introduzione*

Il lavoro presenta i risultati di un'indagine sociofonetica preliminare/esplorativa sul fenomeno residuale della degeminazione di /r:/ nella varietà toscana pratese, attraverso il riuso di un archivio orale storico, inizialmente raccolto per altri scopi.

Le rotiche (Spreafico, Vietti, 2013) sono dal punto di vista fonetico caratterizzate da un'ampia variabilità riguardo i modi e tipi di articolazione a livello sia intra- che interlinguistico (Hall, 1997; Solé, 2002; van Hout, Van de Velde, 2001; Wiese, 2001-2010). Di conseguenza, le rotiche sono difficilmente definibili o classificabili come una classe univoca di suoni (come, ad esempio, quelli occlusivi). Inoltre, rispetto ad altre, la letteratura fonetica a riguardo è più scarsa (Maddieson, 1984; Lindau, 1985; Ladefoged, Maddieson, 1996; Catford, 2001; Foulkes, Docherty, 2001; Scobbie, 2006; Chabot, 2019). Nel sistema consonantico italiano i contesti scempio e geminato (in posizione intervocalica) si distinguono per il tratto di lunghezza temporale della consonante. Nel caso delle rotiche, per la distinzione dei due contesti, sono solitamente osservati sia i parametri di durata della vocale precedente e della rotica (Bertinetto, Vivalda, 1978), che il tipo: *tap*, *trill*, *fricativa*, *approssimante* (Canepari, 1980; Vietti, Spreafico & Romano, 2010; Romano, 2013; Celata, Meluzzi, Ricci, 2016; Celata, Vietti & Spreafico, 2019).

Studi recenti hanno messo in luce come il fenomeno di degeminazione sia sempre meno comune sia a nord che a sud dell'isoglossa La Spezia-Rimini, a causa della progressiva standardizzazione della lingua (Mairano, De Iacovo, 2020). Nelle zone dell'Italia centrale il fenomeno ha riguardato alcune consonanti, tra cui, appunto, le rotiche (Rohlfs, 1966; Giannelli, 1976; Loporcaro, 2009; Nodari, Meluzzi, 2020). In Toscana, la degeminazione si configura come aspetto residuale, derivante dal percolamento di questo tratto dai dialetti settentrionali verso le zone a sud dell'isoglossa (Rohlfs, 1966; Giannelli, 1976; Giannelli, 1983). Tuttavia, il fenomeno risulta essere scarsamente indagato sia dal punto vista della distribuzione areale (Giannelli, 1983) che sociofonetico; pertanto, ben si presta a una verifica su archivi orali del passato. I contributi che hanno recentemente indagato questo fenomeno sono Celata *et al.*, (2019) in due città ad ovest della regione (Pisa e Livorno) e Nodari e Calamai (2021) nell'isola d'Elba.

Questo lavoro si concentra sull'area rurale fiorentina-pratese (Giannelli, 1976) che dal punto di vista geolinguistico, si trova appena al di sotto del confine con l'isoglossa attraverso cui si è diffuso il fenomeno di degeminazione e che, dal punto di vista storico, presenta poche documentazioni linguistiche sulla varietà locale, avendo subito per diversi secoli una forte influenza del fiorentino (Fiorelli, 1980). Secondo quanto riportato in Fiorelli (1980) tra il XIII e il XIV secolo, Prato possedeva una propria individualità come varietà ed era nota per condividere molti tratti linguistici con il fiorentino rispetto altre varietà della Toscana occidentale. A partire dalla seconda metà del XIV secolo, tuttavia, la vicinanza geografica delle due province e l'espansione politica fiorentina portarono all'annessione di Prato all'amministrazione fiorentina. La somiglianza linguistica tra le due province portò il fiorentino ad essere recepito precocemente dai cittadini pratesi, diventando, a partire dalla seconda metà del XIV secolo, la lingua maggiormente utilizzata a Prato, non solo per l'espressione letteraria, ma anche nella comunicazione quotidiana.

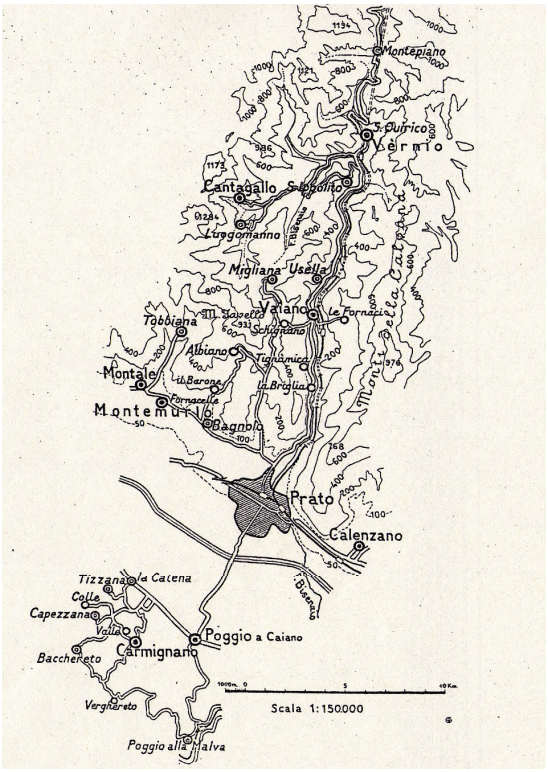
L'unica fonte di dati orali e di tipo spontaneo che permetta di compiere ad oggi un'indagine sulla presenza di degeminazione nella varietà pratese è l'archivio storico orale di Angela Spinelli. Questo lavoro propone di indagare la presenza del fenomeno di degeminazione di /r:/ a livello lessicale in posizione intervocalica in parole bisillabi fonologicamente geminate, oltre che la presenza di correlazioni con variabili di tipo sociale ricavate dalle registrazioni e dal ricco materiale di corredo dell'archivio. La decisione di riutilizzare i dati orali che compongono questo archivio è, inoltre, motivata dalla struttura delle interviste condotte da Angela Spinelli che, avendo come obiettivo la raccolta di testimonianze sul periodo storico delle due guerre mondiali, ha favorito la presenza di un linguaggio poco controllato e più spontaneo da parte dei soggetti intervistati (Labov, 1966-1972), sempre molto difficile da elicitare in laboratorio.

L'articolo è organizzato come segue: la sezione 2 presenta l'archivio di Angela Spinelli, mentre la sezione 3 descrive il campione analizzato e la metodologia adottata per il trattamento dei dati. La sezione 4 presenta le analisi qualitativa e quantitativa. Infine, la sezione 5 discute i risultati, mentre la sezione 6 riporta le conclusioni, i limiti di questa ricerca e le prospettive future.

2. *L'archivio*

L'archivio è stato raccolto dalla storica Angela Spinelli all'inizio degli anni '80. La ricercatrice ha condotto una serie di interviste faccia a faccia con gli abitanti di piccoli paesi della zona rurale di Prato lungo l'alta valle del fiume Bisenzio (Cantagallo, Vaiano, Montemurlo) (Fig. 1) per garantire la conservazione delle memorie sugli eventi storici accaduti durante il periodo post Seconda Guerra Mondiale (Spinelli, 1981). L'archivio è composto da 59 audiocassette (corrispondenti a oltre 120 ore di registrazione) e da un ampio insieme di materiale di corredo, che comprendono importanti informazioni sul protocollo di ricerca ed annotazioni fatte dalla ricercatrice durante il suo lavoro (Fig. 2). L'archivio è stato acquisito legalmente da Angela Spinelli e digitalizzato nel 2011 nell'ambito del progetto Gra.fo. (Calamai, Billotti, 2017). Poiché l'archivio è attualmente non disponibile per la consultazione online dal sito web del progetto, l'accesso ai dati d'archivio è stato possibile tramite la piattaforma Archivio Vi.Vo. (Calamai, Pretto, Stamuli, Piccardi, Candeo, Bianchi & Monachini, 2021) parte dell'infrastruttura italiana di CLARIN-IT[1], sviluppata per garantire la conservazione e il riuso di archivi orali storici per vari scopi (ricerca, disseminazione, valorizzazione ecc.) in sicurezza limitandone la perdita, la frammentazione e il danneggiamento.

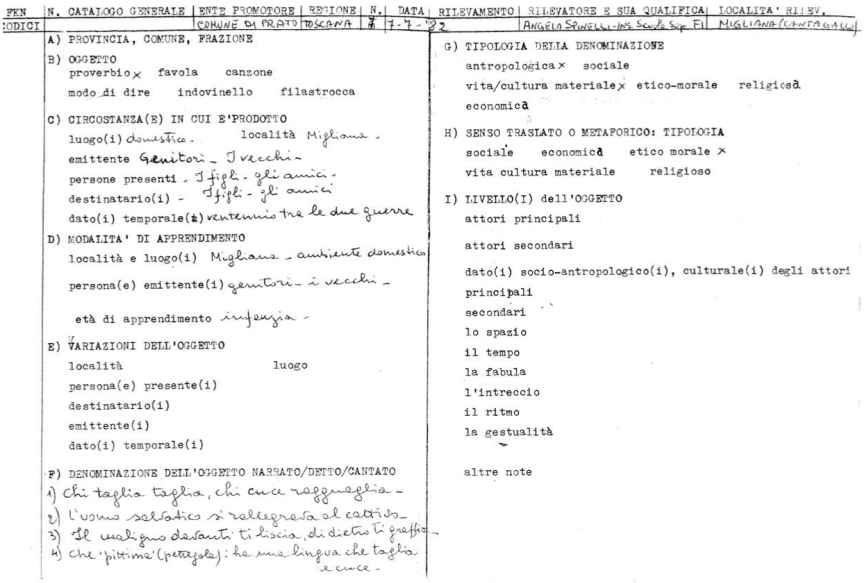
Figura 1 - *Mappa dell'area pratese (Spinelli 1981: 13)*



[1] https://www.clarin-it.it

Angela Spinelli ha condotto la sua ricerca sotto la supervisione del professor Roger Absalom (Absalom, Spinelli, 1998). Il metodo di indagine impiegato dalla ricercatrice appartiene al campo della storia orale che consiste nell'elicitare storie di vita inedite di persone nate nella zona rurale di interesse tra il 1880 e il 1930 e appartenenti a diversi gruppi sociali (ad esempio, piccoli proprietari, mezzadri, affittuari, carbonai, commercianti di carbone, pastori). Lo scopo della ricerca era di raccogliere quante più informazioni possibili riguardo gli aspetti della vita rurale, materiale e associativa che caratterizzava quelle zone nel periodo tra le due guerre mondiali. Inoltre, voleva verificare l'essenza della mentalità contadina a lungo termine[2], che all'epoca era stata messa alla prova dal modello di sviluppo delle piccole e medie industrie. Per fare ciò, Angela Spinelli ha proceduto compiendo un'analisi analizzando i periodi di crisi (guerra) e di transizione (dopoguerra) in cui la vecchia mentalità (rurale) veniva messa in discussione rispetto la nuova (industriale) caratterizzata principalmente da migliori opportunità di lavoro in termini sia di salario che di sforzo fisico (Spinelli, 1988). Un altro aspetto interessante del lavoro di Angela Spinelli è l'uso di parole chiave durante l'interazione con i testimoni per introdurre i vari argomenti di discussione e suscitare la loro memoria su temi come i rifugiati britannici della Seconda Guerra Mondiale ed altri eventi storici (Andreini, Clemente, 2007).

Figura 2 - *Esempio di scheda di corredo dell'archivio Angela Spinelli*


FKN | N. CATALOGO GENERALE | ENTE PROMOTORE | REGIONE | N. | DATA | RILEVAMENTO | RILEVATORE E SUA QUALIFICA | LOCALITA' RILEV.
CODICI | | COMUNE DI PRATO | TOSCANA | 8 | 7-7-82 | | ANGELA SPINELLI-INS. Scuola Sup. FI | MIGLIANA (CONTRADA)

A) PROVINCIA, COMUNE, FRAZIONE

B) OGGETTO
proverbio x favola canzone
modo di dire indovinello filastrocca

C) CIRCOSTANZA(E) IN CUI E'PRODOTTO
luogo(i) domestico - località Migliana -
emittente Genitori - I vecchi -
persone presenti - I figli - gli amici -
destinatario(i) - I figli - gli amici
dato(i) temporale(i) ventennio tra le due guerre

D) MODALITA' DI APPRENDIMENTO
località e luogo(i) Migliana - ambiente domestico
persona(e) emittente(i) genitori - i vecchi -
età di apprendimento infanzia -

E) VARIAZIONI DELL'OGGETTO
località luogo
persona(e) presente(i)
destinatario(i)
emittente(i)
dato(i) temporale(i)

F) DENOMINAZIONE DELL'OGGETTO NARRATO/DETTO/CANTATO
1) Chi taglia taglia, chi cuce raggomeglia -
2) L'uomo salvatico si rallegrerà al cattivo -
3) Il mal'ingno davanti ti lisicia, di dietro ti graffia -
4) Che 'pittima' (pettegola): ha una lingua che taglia e cuce -

G) TIPOLOGIA DELLA DENOMINAZIONE
antropologica x sociale
vita/cultura materiale x etico-morale religiosa
economica

H) SENSO TRASLATO O METAFORICO: TIPOLOGIA
sociale economica etico morale x
vita cultura materiale religioso

I) LIVELLO(I) dell'OGGETTO
attori principali
attori secondari
dato(i) socio-antropologico(i), culturale(i) degli attori principali secondari
lo spazio
il tempo
la fabula
l'intreccio
il ritmo
la gestualità
altre note

I soggetti che hanno partecipato all'indagine sono stati selezionati in base all'età, all'origine e alle raccomandazioni degli altri partecipanti. Il protocollo di intervista

[2] Per mentalità contadina a lungo termine si intende il tipo di mentalità contadina per cui i cambiamenti sono molto più lenti rispetto alla mentalità artigianale e operaia che hanno transizioni più immediate (Spinelli, 1988).

consisteva in tre giorni di colloqui e un giorno di pausa in cui Angela Spinelli ascoltava le registrazioni, riassumeva il suo lavoro e ragionava riguardo a con chi e dove continuare le successive interviste (Andreini, Clemente, 2007). I soggetti selezionati potevano essere intervistati singolarmente o in coppia (ad esempio, madre e figlio, marito e moglie). Le interviste iniziavano con la raccolta di informazioni personali, sulla famiglia e la rete sociale come strategia iniziale per far sì che potessero sentirsi a loro agio. A partire da quel momento, la ricercatrice invitava i soggetti intervistati a parlare di argomenti come: l'alfabetizzazione, la loro relazione con gli alleati britannici, la religione (cerimonie e pellegrinaggi), il fascismo, la condizione femminile (come governanti, infermiere, suore), l'agricoltura (i cicli di lavoro del carbone, del grano, delle castagne, della vendemmia, gli strumenti di lavoro), le malattie, la cultura popolare (sistema proverbiale, leggende, superstizioni) (Fig. 2), la dote e la successione ereditaria femminile, l'emigrazione, l'economia, la struttura familiare, le esperienze di vita durante la prima e la seconda guerra mondiale, i sistemi di matrimonio, il possesso della terra, il periodo di resistenza dopo la Seconda Guerra Mondiale e i fenomeni di mobilitazione per cui le persone iniziarono ad abbandonare la vita rurale per un lavoro migliore e uno stile di vita nelle fabbriche tessili e nei cantieri ferroviari nella vicina città di Prato. Questa descrizione mostra come l'archivio Spinelli sia notevolmente ricco non solo per il numero di ore di registrazione, ma anche per la qualità delle interviste di storia orale; pertanto, ben si presta per scopi di riutilizzo nella ricerca (sociolinguistica e fonetica).

3. *Dati e metodologia*

Il campione di dati orali analizzati in questa ricerca corrisponde a metà archivio Spinelli. Per condurre questa indagine sono state selezionate 31 interviste (su un totale di 62) per un totale di quasi 35 ore di registrazione. Il numero di partecipanti che ha preso parte alle 31 interviste corrisponde a 41, di cui però 3 sono stati esclusi dalle analisi per mancanza di produzioni rilevanti per la ricerca. Per ciascuno dei 38 parlanti sono state raccolte durante l'ascolto delle registrazioni e grazie alla consultazione del materiale di corredo le seguenti informazioni[3]: il sesso, l'anno di nascita, il luogo di residenza, il livello di scolarizzazione, il tipo di impiego e la presenza di esperienze di mobilità (ad esempio, trasferimenti in altri paesi e province per motivi coniugali ed esperienze di sfollamento e di guerra in altre città e stati). I parlanti analizzati stati suddivisi in: sesso maschile e sesso femminile, nati tra il 1890 e il 1933 e provenienti da diverse località afferenti ai comuni di Cantagallo e Vaiano. Il livello di scolarizzazione nel corpus è stato suddiviso in quattro categorie: parlanti che non hanno ricevuto alfabetizzazione, che hanno frequentato le scuole elementari fino ed oltre la terza elementare (livello base e medio), e che hanno approfondito gli studi oltre la quinta elementare (avanzato). Il tipo di impiego è stato suddiviso in maniera dicotomica:

[3] La decisione di convertire le informazioni di nome e cognome in codici univoci composti da iniziali del nome, cognome e sesso (M/F) (per i casi di omonimia è stato aggiunto un numero progressivo) per garantire l'anonimizzazione e riservatezza dei dati è stata approvata dal Comitato Bioetico dell'Università di Pisa in data 01/08/2022.

parlanti con esperienze di lavoro unicamente legate alla terra o altre esperienze oltre al lavoro della terra[4]. Lo stesso criterio è stato utilizzato per l'esperienza di mobilità.

Tabella 1 - *Distribuzione di parlanti, rispetto ai parametri sociali considerati, nel corpus*

Variabili	*Varianti*	*N.*
Sesso	Maschi	20
	Femmine	18
Luogo di residenza	Cantagallo	36
	Vaiano	2
Livello di scolarizzazione	Non alfabetizzati	1
	Base (1/2/3 elementare)	20
	Medio (4/5 elementare)	6
	Avanzato (oltre 5 elementare)	2
	NA	9
Impiego	Terra	26
	Altro	9
	NA	3
Mobilità	Sì	14
	No	19
	NA	5

Per l'estrazione dei dati acustici ed informazioni dall'archivio, si è proceduto inizialmente con l'ascolto delle registrazioni[5] da cui sono state individuate le varie interviste[6]. Successivamente, per ciascuna intervista sono state segmentate e trascritte, con il software ELAN, le parti di discorso rilevanti per l'analisi, ovvero quelle contenenti occorrenze di parole target con /r:/ o /r/ in posizione intervocalica. Di seguito, le sezioni di audio (selezionate in ELAN) sono state importate nel software PRAAT (Boersma, Weenink, 2021) per procedere con l'annotazione fonetica delle parole target seguendo il protocollo proposto da Celata *et al.* (2016: 4). Per mezzo dell'ascolto, ispezione della forma d'onda e dello spettrogramma è stata svolta l'annotazione su tre livelli (*tiers*) (Fig. 3 e 4): nel primo *tier* è stata inserita la parola target (*word*), nel secondo *tier* l'accento della sillaba (*stress*) e nel terzo *tier* la rotica, individuata a partire dai punti di fine della vocale precedente ed inizio della vocale successiva (Esposito, Di Benedetto, 1999: 2053; Ham, 2001: 33).

In questo studio esplorativo, si è deciso di indagare la presenza di degeminazione di /r:/ in posizione intervocalica a livello lessicale limitatamente in parole di tipo bisillabi analizzando solo i valori di durata, (escludendo per il momento l'analisi del tipo di rotica) di: parole fonologicamente geminate che all'ascolto dell'autrice sono risultate prodotte come geminate (terra > terra) (107); parole fonologicamente

[4] Nelle tabelle etichettati rispettivamente come "terra" e "altro".

[5] Digitalizzate e disponibili in formato .wav.

[6] Che corrispondono alle unità documentali. Per maggiori informazioni, si faccia riferimento a Calamai *et al.* (2021) e Luzietti (in pubblicazione).

geminate che all'ascolto sono risultate prodotte come scempie (terra > tera) (157); e parole fonologicamente scempie prodotte come scempie (sera) (20). Sulla base di questo criterio le parole analizzate sono: terra, terre, guerra, ferri, ferro, sera, serra, carro, carri, per un totale di 284 occorrenze (Tab. 2, 3).

Figura 3 - *Esempio di annotazione della parola "terra"*

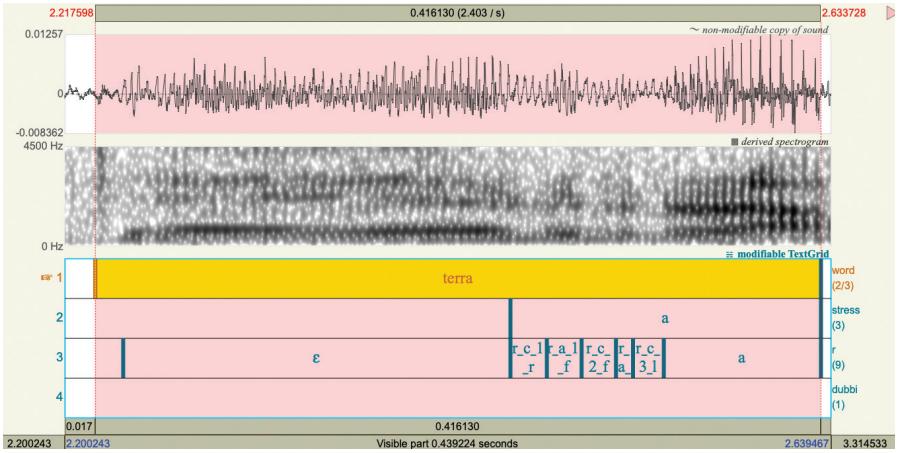
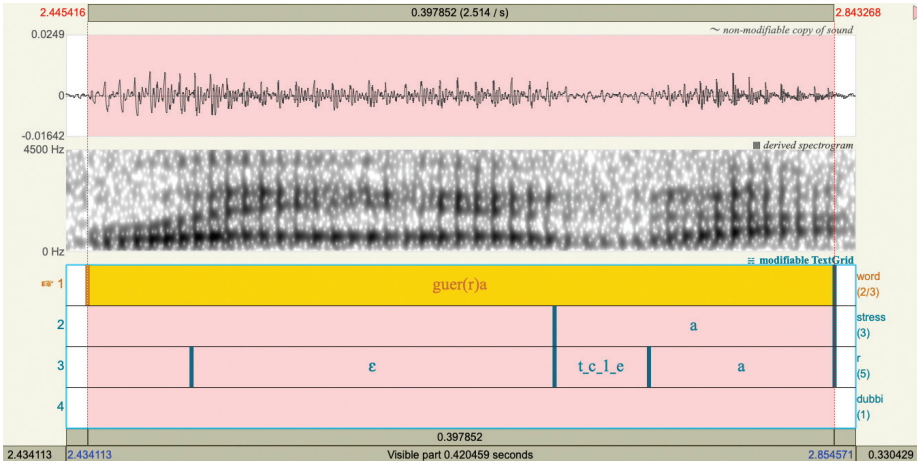


Figura 4 - *Esempio di annotazione della parola "guerra" udita come "guera"*



4. *Analisi*

In questa sezione sono presentati i risultati delle analisi qualitativa e quantitativa.

4.1 Analisi qualitativa

Come specificato nel paragrafo precedente, la presenza di degeminazione è stata determinata su base uditiva dall'autrice nel corso dell'annotazione dei dati. Le

seguenti Tab. 2, 3 e 4 servono a comprendere come la classificazione (percepita come scempia/percepita come geminata) su base uditiva si ponga rispetto all'effettiva durata delle vocali precedenti, rotiche e V:VC *ratio*.

Da una prima osservazione delle informazioni Tab. 2 e 3, si nota che, coerentemente con il tema delle interviste nell'archivio, la parola "guerra" presenta il maggior numero di occorrenze nel corpus ed è più frequentemente percepita dall'autrice come scempia. Lo stesso vale per la parola "terra". Confrontando poi i valori medi di durata delle vocali precedenti (VP) e rotiche, sempre in Tab. 2 e 3, si può osservare che, in linea con quanto affermato da Bertinetto e Vivalda (1978), nelle parole etichettate come scempie rispetto a quelle geminate vi è incremento della durata della vocale precedente e riduzione della durata della rotica nelle parole "guerra", "terra", "ferri" e "ferro" (presenti in entrambe le tabelle). Nel lavoro di Ham (2001), basatosi a sua volta sul lavoro di Kohler (1977), viene riportato che i valori di V:VC *ratio* oltre 0,70 indicano la presenza di un allungamento della vocale precedente con accorciamento della rotica. Viceversa, valori di V:VC *ratio* inferiori a 0,60 indicherebbero un accorciamento della durata della vocale precedente con conseguente allungamento della durata della rotica. Osservando i valori di V:VC *ratio* delle parole in Tab. 2 e 3, per le parole etichettate come scempie, i valori corrispondono a >0,70. Al contrario, i valori di V:VC *ratio* per le parole etichettate come geminate, i valori non risultano essere <0,60, bensì oscillano tra 0,60 e 0,70.

Tabella 2 - *Totale occorrenze, valori medi di durata della vocale precedente (VP), rotica e V:VC ratio di parole fonologicamente geminate /r:/ uditivamente percepite come geminate /r:/*

Parole /r:/	*Totale occorrenze*	*Uditivamente geminate /r:/*	*Media durata VP (ms)*	*Media durata /r:/ (ms)*	*V:VC ratio*
guerra	125	47	144	70	0,67
terra	65	29	182	91	0,67
ferri	7	3	174	113	0,61
ferro	4	3	134	57	0,70
carri	1	1	293	183	0,62
carro	1	1	189	27	0,87
serra	1	1	165	82	0,67
terre	2	1	NA	NA	NA

Tabella 3 - *Totale occorrenze, valori medi di durata della vocale precedente (VP), rotica e V:VC ratio di parole fonologicamente geminate /r:/ uditivamente percepite come scempie /r/*

Parole /r:/	*Totale occorrenze*	*Uditivamente scempie /r/*	*Media durata VP (ms)*	*Media durata /r/ (ms)*	*V:VC ratio*
guerra	125	78	153	41	0,79
terra	65	36	203	41	0,83
ferri	7	4	147	42	0,78
ferro	4	1	71	29	0,71
terre	2	1	216	17	0,93

Nell'ultimo confronto tra i valori riportati in Tab. 3 e quelli relativi alla parola "sera" in Tab. 4, emerge solamente una più netta differenza in termini di durata con la parola "ferro" (che, tuttavia, ha una sola occorrenza). Il confronto con i restanti valori di durata delle parole percepite come uditivamente scempie sembra essere in linea con i valori di "sera", parola considerata di riferimento in questo lavoro.

Tabella 4 - *Numero di occorrenze, valore medo di durata della vocale precedente (VP), valori medi di durata della rotica e V:VC ratio della parola sera fonologicamente scempia /r/ uditivamente scempia /r/*

Parola /r/	*Totale occorrenze*	*Uditivamente scempie /r/*	*Media durata VP (ms)*	*Media durata /r/ (ms)*	*V:VC ratio*
sera	19	19	139	29	0,83

Di seguito, in Tab. 5, 6, 7 e 8, sono riportate informazioni riguardo ai soggetti intervistati da Angela Spinelli suddivisi in base al sesso per cui è sempre o non è mai stata percepita degeminazione di /r:/.[7]

Tabella 5 - *Parlanti FEMMINE in cui è SEMPRE udita degeminazione*

Parlante	*Anno di nascita*	*Luogo di residenza*	*Impiego*	*Esperienza di mobilità*	*Livello di Scolarizzazione*
CPF	1919	Migliana	Terra	Sì	3
EPF	1904	Migliana	Terra	No	NA
FGF	1890	Castello	Terra	Sì	3
GGF	1910	Parugiano	Terra	No	3
LFF	NA	NA	NA	NA	NA
MBF2	1927	Migliana	Altro	Sì	3
RPF	1906	Migliana	Terra	No	1

Tabella 6 - *Parlanti FEMMINE in cui non è MAI udita degeminazione*

Parlante	*Anno di nascita*	*Luogo diresidenza*	*Impiego*	*Esperienza di mobilità*	*Livello di Scolarizzazione*
AAF	1925	Castello	Terra	Sì	2
FMF	NA	NA	NA	NA	NA
LVF	1921	Migliana	Altro	Sì	2
MPF	1893	Riottolo	Terra	No	3

Dal confronto tra parlanti di sesso femminile e maschile rispetto alla presenza o assenza di degeminazione, non emergono informazioni rilevanti che possano essere ricondotte ai parametri sociali dei parlanti nei due gruppi. L'unica osservazione possibile, al momento, è che la distribuzione tra coloro che degeminano sempre (7), non degeminano mai (4) ed alternano (7) nel gruppo di parlanti di sesso femminile

[7] Dai risultati in Tab. 5, 6, 7 e 8 sono state escluse tutte le occorrenze della parola "sera".

è meno netta rispetto a coloro che degeminano sempre (6), non degeminano mai (1) ed alternano (13) di sesso maschile.

Tabella 7 - *Parlanti MASCHI in cui è SEMPRE udita degeminazione*

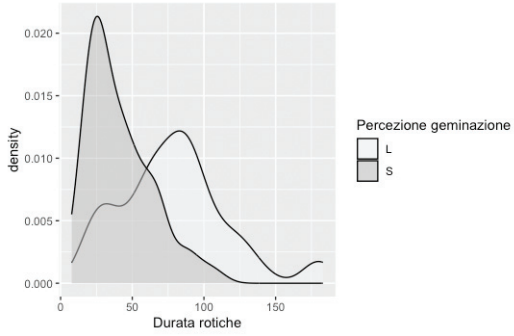
Parlante	*Anno di nascita*	*Luogo di residenza*	*Impiego*	*Esperienza di mobilità*	*Livello di Scolarizzazione*
GGM2	1913	Castello	Terra	Sì	3
GGM	1933	Luicciana	Altro	Sì	4
GMM	1933	NA	Terra	NA	NA
GPM0	1929	Pratalecchio	Terra	No	4
NUDPM	1920	Masseto	Terra	No	4
UMM	NA	NA	NA	NA	NA

Tabella 8 - *Parlanti MASCHI in cui non è MAI udita degeminazione*

Parlante	*Anno di nascita*	*Luogo di residenza*	*Impiego*	*Esperienza di mobilità*	*Livello di Scolarizzazione*
DPM	1921	Migliana	Terra	No	5

Al fine di esaminare la densità di distribuzione della durata delle rotiche percepite come geminate e scempie, è stata creata la Figura 5.

Figura 5 - *Densità di distribuzione dei valori di durata delle rotiche divise in base al tipo di percezione: geminata (L) o scempia (S)*

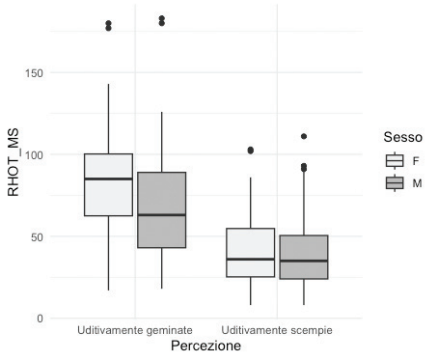


L'asse x rappresenta i valori di durata delle rotiche (ms), mentre l'asse y rappresenta la densità dei dati. L'altezza della curva riflette la probabilità che i dati si trovino in una determinata regione, mentre la presenza di picchi nella curva indica dove i dati sono maggiormente concentrati. In questo caso, la curva delle parole percepite come scempie (etichettate nella matrice come "S") presenta un solo picco notevolmente più alto rispetto a quello della curva di distribuzione delle parole percepite come geminate (etichettate nella matrice come "L"). Pertanto, il confronto tra questi due gruppi presenta, come atteso, che i valori di durata delle parole percepite come

scempie (S) siano più simili tra loro, mentre vi è una maggiore variabilità nella durata delle rotiche percepite come geminate (L).

La presenza di maggiore variabilità nei valori di durata delle rotiche percepite come geminate è ulteriormente approfondita nel *boxplot* sottostante (Figura 6).

Figura 6 - *Distribuzione dei valori di durata delle rotiche tra parlanti di sesso maschile e femminile*



Dal confronto tra parlanti di sesso maschile e femminile, emerge che la durata delle rotiche percepite come scempie è simile in entrambi i gruppi. Al contrario, la presenza di differenza di durata tra le rotiche percepite come geminate tra i due gruppo, dove le femmine mostrano una maggiore durata, suggerendo la presenza di iperarticolazione.

4.2 Analisi quantitativa

Per l'analisi statistica, sono stati utilizzati i modelli di regressione lineare a effetti misti in R (versione 4.3.0) e il pacchetto lme4 (Bates, Maechler, Bolker & Walker, 2015). Il dataset è composto da 170 token[8]. Per l'analisi sono state selezionate due variabili dipendenti: (i) durata della consonate rotica e (ii) durata della vocale precedente. Per ogni variabile sono stati eseguiti modelli misti lineari (Bates *et al.*, 2015) includendo sei diverse variabili fisse: (i) sesso dei parlanti (maschio/femmina); (ii) la percezione di degeminazione[9]; (iii) il tipo di impiego (terra/altro); (iv) l'esperienza di mobilità (sì/no), (v) l'anno di nascita (variabile continua); (vi) il livello di scolarizzazione (variabile categorica a quattro livelli, analfabeta, base, medio, avanzato) e (vii) la durata in ms. della parola contenente la rotica in esame, messa a fattore per poter avere una idea anche della velocità d'eloquio dei diversi locutori. I parlanti e le parole sono stati inseriti come fattore *random*. Per trovare il modello che meglio si adattasse ai dati è stata utilizzata la funzione *step* del pacchetto lmerTest (Kuznetsova, Brockhoff & Christensen,

[8] Al netto dei valori di durata e variabili sociali indicati con NA e la parola "sera".

[9] Le rotiche fonologicamente geminate udite come geminate sono state etichettate nella matrice con "L". Al contrario, quelle udite come scempie con "S"

2017) che esegue l'eliminazione automatica di tutti gli effetti non significativi confrontando gli *Akaike Information Criterion* (AIC) derivanti dall'esclusione di ogni variabile candidata e dall'aggiunta di ogni variabile candidata tra gli insiemi di regressione superiori e inferiori forniti dal modello, nonché dall'esclusione o dall'aggiunta della variabile che fornisce il miglior AIC.[10] I risultati dei modelli misti lineari che meglio si sono adattati alla durata della rotica e vocale precedente sono riportati rispettivamente in Tab. 9, 10 e 11.

Tabella 9 - *Risultati del modello misto lineare di miglior adattamento per la durata delle rotiche; numero di osservazioni: 170, soggetti: 29, parole: 6*

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)[11]	-4.50811	14.63435	170.00000	-0.308	0.75842
Sesso maschile	-10.82466	3.96145	170.00000	-2.732	0.00695**
Scolarizzazione - base	11.36839	9.74434	170.00000	1.167	0.24498
Scolarizzazione - medio	8.80640	10.14336	170.00000	0.868	0.38651
Scolarizzazione - avanzato	45.93121	14.32417	170.00000	3.207	0.00161**
Impiego: terra	14.08262	7.74548	170.00000	1.818	0.07080.
Mobilità: sì	6.30405	9.90867	170.00000	0.636	0.52549
Percezione degeminazione	-33.12269	4.35981	170.00000	-7.597	1.93e-12***
Durata parola	0.16705	0.01728	170.00000	9.668	< 2e-16***
Impiego: terra & Mobilità: sì	-33.55382	10.88814	170.00000	-3.082	0.00240**
Mobilità: sì & percezione degeminazione	18.90727	8.91312	170.00000	2.121	0.03535*

I risultati del modello misto lineare mostrano la presenza di differenze significative ($p < 0,05$) nella durata delle consonanti rotiche. In particolare, si osserva un accorciamento significativo nelle rotiche prodotte da parlanti di sesso maschile e in parole fonologicamente geminate percepite come scempie. Al contrario, i parlanti con livelli di scolarizzazione avanzati (oltre la quinta elementare) tendono a produrre le rotiche in maniera significativamente più lunga rispetto a chi è meno o non è

[10] Partendo da un modello iniziale che include tutte le variabili, l'eliminazione di tutti gli effetti non significativi avviene una variabile alla volta. L'eliminazione della parte casuale viene eseguita per prima, seguita dall'eliminazione della parte fissa. L'eliminazione della parte casuale viene effettuata utilizzando il test del rapporto di verosimiglianza; se è presente una correlazione tra la pendenza e l'intercetta, il modello semplificato manterrà solo l'intercetta.

[11] I valori di rifermento per l'intercetta sono: "sesso: femminile, impiego: altro, mobilità: no, percezione geminazione".

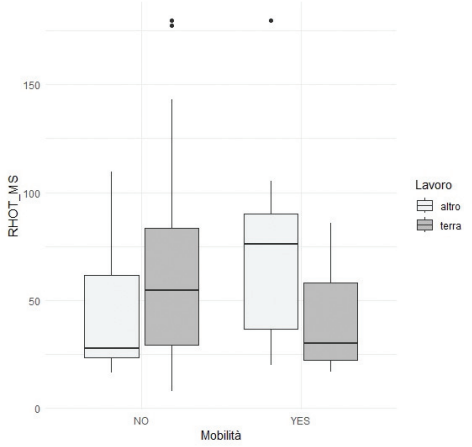
scolarizzato. Sono state inoltre riscontrate diverse interazioni significative tra mobilità e lavoro (Figura 7) e tra mobilità e percezione di degeminazione (Figura 8). Al pari del tipo di impiego l'esperienza di mobilità sembrerebbe in grado di influenzare la presenza o assenza di degeminazione nel campione. Infine, emerge una correlazione positiva tra durata della rotica e della parola.

In linea con quanto riportato in Tab. 9 sono riportati di seguito i risultati dell'analisi chi-quadro per mezzo del pacchetto *afex* (Winter, 2019) con riferimento alle variabili inserite nel modello per l'analisi che rivelano diverse associazioni significative.

Tabella 10 - *Risultati di analisi del chi-quadro - df full model: 14*

	Effect	*df*	*Chisq*	*p. value*
1	*Sesso*	1	6.65 **	.010
2	*Livello di scolarizzazione*	3	12.52 **	.006
3	*Impiego*	1	0.22	.637
4	*Mobilità*	1	0.04	.846
5	*Percezione*	1	26.69 ***	<.001
6	*Durata parola*	1	74.22***	<.001
7	*Impiego: mobilità*	1	9.24 **	.002
8	*Mobilità: percezione*	1	4.44*	.035

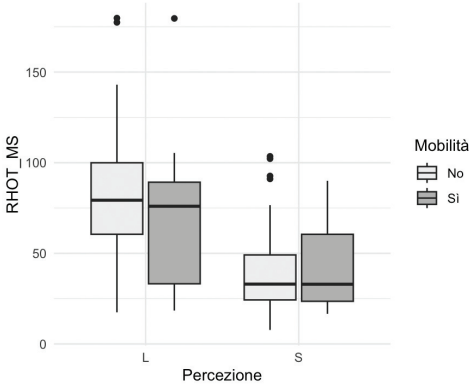
Figura 7 - *Distribuzione dei valori di durata delle rotiche rispetto alle esperienze di mobilità nei parlanti che hanno dichiarato come impiego il solo lavoro della terra e altri lavori*



Esaminando più attentamente l'interazione tra l'esperienza di mobilità e il tipo di impiego nel campione, emergono due pattern opposti. Come illustrato nella Figura 7, si nota un accorciamento significativo tra coloro che non hanno avuto esperienza di mobilità e hanno svolto lavori diversi da quelli legati alla terra. Tuttavia, tale accorciamento non è osservato tra coloro che hanno avuto esperienza di mobilità ma hanno svolto solo lavori legati alla terra. Al contrario, nel gruppo di coloro che hanno avuto esperienza di

mobilità, si rileva un accorciamento significativo tra coloro che hanno svolto solo lavori legati alla terra, ma non tra coloro che hanno svolto altri tipi di lavoro.

Figura 8 - *Distribuzione dei valori di durata delle rotiche rispetto alle esperienze di mobilità nei parlanti e di percezione di geminazione e degeminazione in parole fonologicamente geminate*



Rispetto poi all'interazione tra l'esperienza di mobilità e la percezione di degeminazione risultata significativa nel modello, nella Figura 8 si può notare come al pari della presenza di parole percepite come degeminate (S) che mostrano valori di durata effettivamente più brevi in entrambi i gruppi di parlanti (mobili e non mobili), i valori di durata di parlanti con esperienze di mobilità presentano maggiore variabilità, lasciando presupporre la possibilità che ciò sia dato dall'influenza di altre varietà con cui sono stati in contatto in ambito lavorativo o durante la guerra.

Tabella 11 - *Risultati del modello misto lineare di miglior adattamento per la durata della vocale precedente; numero di osservazioni: 170, soggetti: 29, parole: 6*

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)[12]	0.34783	23.75903	42.35975	0.015	0.988
Sesso maschile	-2.26275	6.82003	17.91935	-0.332	0.744
Scolarizzazione - base	1.06078	16.57038	23.73302	0.064	0.949
Scolarizzazione - medio	-3.03941	16.99310	23.52421	-0.179	0.860
Scolarizzazione - avanzato	-3.81637	23.32544	33.75911	-0.164	0.871
Impiego: terra	6.31302	12.20912	32.08249	0.517	0.609
Mobilità: sì	0.90629	15.19049	48.79577	0.060	0.953
Percezione degeminazione	25.39981	6.28072	141.10777	4.044	8.61e-05 ***

[12] I valori di rifermento per l'intercetta sono: "sesso: femminile, impiego: altro, mobilità: no, percezione geminazione".

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
Velocità parola	0.38849	0.02477	165.99422	15.684	< 2e-16 ***
Impiego: terra & Mobilità: sì	11.09453	16.75977	44.64351	0.662	0.511
Mobilità: sì & percezione degeminazione	-9.99703	12.64168	145.92410	-0.791	0.430

L'analisi dei risultati del modello misto lineare (Tab. 11) rivela differenze significative ($p < 0,05$) nella durata delle vocali precedenti. In particolare, si osserva un significativo allungamento nella durata delle vocali precedenti nelle parole percepite come degeminate, a sostegno dell'ipotesi della presenza di scempiamento nel campione. È altresì rilevante notare la presenza di significatività nella correlazione positiva tra la durata della vocale precedente e la velocità della parola per cui si osserva che parole più lunghe conducono a un allungamento sia delle vocali che delle rotiche.

5. *Discussione*

L'obiettivo di questo studio è stato quello di verificare la presenza di degeminazione di /r:/ in posizione intervocalica a livello lessicale in parole bisillabi nell'area pratese e comprendere se tale fenomeno fosse correlato a determinate/ad alcune variabili sociali. Attraverso l'ascolto delle registrazioni delle interviste contenute nell'archivio Spinelli, sono state identificate una serie di parole target per l'analisi. Durante la fase di ascolto, le parole target sono state classificate in percepite come uditivamente scempie o geminate e poi annotate. A partire da questa classificazione, nella prima parte dell'analisi qualitativa, sono stati osservati i valori di durata delle vocali precedenti e rotiche nelle parole "guerra", "terra", "ferri" e "ferro" che hanno mostrato essere coerenti con le aspettative in Bertinetto e Vivalda (1978). Al contrario, i valori derivanti dal calcolo del V:VC *ratio* non hanno mostrato una differenza altrettanto netta e coerente con le aspettative. In linea con i valori del V:VC *ratio* calcolati nelle Tab. 2 e 3, la Figura 5 nella quale viene rappresentata la densità dei valori di durata delle rotiche percepite come geminate o scempie, mostra come vi sia effettivamente una maggiore concentrazione (e quindi maggiore similarità) tra i valori di durata delle rotiche prodotte come scempie (con V:VC *ratio* in linea con le aspettative) e una maggiore variabilità nei valori di durata delle rotiche prodotte come geminate (con V:VC *ratio* meno in linea con le aspettative). In presenza di ulteriori test percettivi, oltre che un maggior numero di dati, si provvederà ad approfondire la comprensione dei risultati e identificare altri fattori che potrebbero aver potenzialmente influenzato questa analisi, come, ad esempio, la poca familiarità dell'autrice con la varietà linguistica in esame.

Successivamente, è stato effettuato un confronto tra le produzioni dei soggetti intervistati in base al sesso e alla (percepita) totale presenza o assenza

di degeminazione di /r:/. Le Tab. 5, 6, 7 e 8 hanno evidenziato come nel gruppo maschile prevalga l'alternanza di parole (fonologicamente geminate) prodotte sia come geminate che come scempie. D'altro canto, il gruppo femminile mostra una distribuzione più equilibrata tra coloro che producono le parole target come geminate, scempie o in alternanza. È interessante notare che, nonostante ciò, dalla Figura 6 si osserva una maggiore somiglianza nella durata delle rotiche tra maschi e femmine nelle parole prodotte come scempie e una maggiore variabilità nelle parole prodotte come geminate. Per un'analisi più approfondita (dell'atteggiamento linguistico di questi parlanti), sarà fondamentale prendere in considerazione il numero di occorrenze di geminazione e scempiamento di /r:/ da confrontare con coloro che presentano entrambi i casi.

L'analisi quantitativa ha evidenziato risultati degni di nota seppur molto preliminari. Trattandosi, infatti, di dati storici provenienti da un archivio orale, a differenza dei casi in cui i dati vengono raccolti appositamente per eseguire questo tipo di esperimenti, si sono riscontrati dati e/o variabili mancanti in alcune circostanze. Ciò ha generato uno sbilanciamento nel corpus e una riduzione dei valori osservabili (si veda § 4.2), passando da 284 a 170. Ad ogni modo, l'analisi mediante modelli misti di regressione fornisce comunque una, seppur preliminare, panoramica sulla presenza del fenomeno di degeminazione nell'archivio Spinelli. Questi primi risultati in merito alla presenza di accorciamento della durata della rotica nei parlanti di sesso maschile, all'allungamento nei soggetti con un livello di scolarizzazione più alto e l'allungamento della durata della vocale precedente nelle parole percepite come degeminate, forniscono importanti spunti di riflessione. Da un lato, in prospettiva sociolinguistica, apre la possibilità di svolgere indagini in merito al rapporto tra sesso dei parlanti di area toscana e l'uso di forme più o meno (percepite come) prestigiose (Trudgill, 1972: 174-194). Dall'altro, dal punto di vista fonetico, permette di esplorare ulteriormente la presenza di allungamento temporale della vocale precedente per i casi in cui è presente scempiamento della consonante rotica (Bertinetto, Vivalda, 1978).

6. *Conclusione*

L'articolo presenta i risultati di un'indagine sociofonetica preliminare sul fenomeno residuale della degeminazione di /r:/ nella varietà toscana pratese, attraverso il riuso dell'archivio storico orale raccolto dalla ricercatrice Angela Spinelli, inizialmente raccolto per altri scopi. L'interesse per questo lavoro nasce dal fatto che in passato, alcuni autori hanno segnalato la presenza di fenomeni di scempiamento di rotiche in territori vicini indagati dalla storica. Tuttavia, ad eccezione di due recenti lavori (Celata *et al.*, 2019; Nodari, Calamai, 2021) il fenomeno è stato poco indagato dal punto di vista sperimentale.

Dal punto di vista della raccolta dati, trattandosi di un fenomeno fonetico residuale, la tipologia di narrazioni (spontanee) raccolte nell'archivio Spinelli non sono ad oggi facilmente elicitabili in laboratorio, tanto meno la tipologia

di parlanti intervistati. Dal punto di vista dell'analisi sperimentale sono emerse difficoltà dal punto di vista procedurale che hanno reso il lavoro molto oneroso come, ad esempio: la scarsa qualità del segnale, la non equa distribuzione delle ricorrenze (Tab. 2) e la mancanza di alcune informazioni (Tab. 3, 4, 5, 6). Tuttavia, questo lavoro presenta sfide di grande portata in quanto permette di ragionare sui limiti del riuso dei dati orali, sul rapporto tra tempo reale e tempo apparente in sociolinguistica e sullo stato del fenomeno fonetico nell'epoca di raccolta dei dati dell'archivio.

Rispetto ai risultati ottenuti in questa fase preliminare del lavoro, è possibile presumere la presenza di degeminazione di /r:/ tra i parlanti che hanno preso parte all'indagine di Angela Spinelli, si estenda al resto dell'archivio. Sulla base dei risultati ottenuti, si delineano diverse prospettive di ricerca future. In primo luogo, l'ampliamento del corpus al fine di ottenere una rappresentazione completa del fenomeno all'interno dei dati d'archivio. In secondo luogo, la realizzazione di analisi percettive più oggettive e mirate all'identificazione dei fenomeni di degeminazione da mettere a fattore nel lavoro di analisi tramite modelli. Al contempo, si provvederà anche a verificare la presenza di fenomeni di accomodamento da parte dei parlanti nei confronti della storica nel corso delle varie interviste.

Ringraziamenti

Un ringraziamento speciale a Rosalba Nodari, Duccio Piccardi e Riccardo Del Gratta per il loro prezioso aiuto durante la realizzazione di questo lavoro, in particolare per l'analisi statistica.

Riferimenti bibliografici

Absalom, R. Spinelli, A. (1998). Contadini e gli ex prigionieri anglo-sassoni a prato nella seconda guerra mondiale. In Becattini, G. (a cura di), *Prato Storia di una città. Il modello pratese (1943 – ad oggi)*. Firenze, Le Monnier IV, 43–81.

Andreini, A. & Clemente, P. (2007). *I custodi delle voci. Archivi orali in Toscana: primo censimento* (Vol. 8), Firenze, Italy: Regione Toscana.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. In *Journal of Statistical Software*, 67, 1, 1-48.

Bertinetto, P.M., Vivalda, E. (1978). Recherches sur la perception des opposition de quantité en italien. In *Journal of Italian Linguistics*, 3, 97-116.

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [software] Versione 6.3.08. https://www.praat.org/

Calamai, S., Biliotti, F. (2017). The Gra.fo project: from collection to dissemination. In *Umanistica Digitale*, 1:85–103.

Calamai, S., Bertinetto, P.M. (a cura di) (2014). *Le soffitte della voce. Il progetto Grammo-foni.* Manziana, Italy: Vecchiarelli.

Calamai, S., Pretto, N., Stamuli, M.F., Piccardi, D., Candeo, G., Bianchi, S., & Monachini, M. (2021). Community-Based Survey and Oral Archive Infrastructure in the Archivio Vi. Vo. Project. In *Selected papers from the CLARIN Annual Conference 2020*. Linköping: Linköping University Electronic Press. 180: 55-64.

Calamai, S., Nodari, R. (2022). Fare interviste, provocare lingue: l'esperienza della carta dei dialetti italiani. In *Imparare dagli errori. Difficoltà, complicazioni, ripensamenti nella storia orale.* Convegno AISO Treviso/Venezia, 10-11 ottobre 2022.

Canepari, L. (1980). *Italiano standard e pronunce regionali*. Cleup.

Catford, J.C. (2001). On Rs, rhotacism and paleophony. In *Journal of the International Phonetic Association,* 31(2). 171–185.

Chabot, A. (2019). What's wrong with being a rhotic? In *Glossa: A Journal of General Linguistics,* 4(1): 38. 1–24. https://doi.org/10.5334/gjgl.618

Celata, C., Meluzzi, C. & Ricci, I. (2016). The sociophonetics of rhotic variation in Sicilian dialects and Sicilian Italian: Corpus, methodology and first results. In *Loquens*, 3, 1, e025.

Celata, C., Vietti, A. & Spreafico, L. (2019). An articulatory account of rhotic variation in Tuscan Italian: Synchronized UTI and EPG data. In *Romance Phonetics and Phonology*. In Gibson, M. and Gil, J. (a cura di) Oxford University Press, Oxford, 92-117.

Esposito, A., Di Benedetto, M.G. (1999). Acoustical and perceptual study of gemination in Italian stops. In *Journal of the Acoustical Society of America*, 106, 2051-2062.

Fiorelli, P. (1980). Il linguaggio dei pratesi. In *Storia di Prato*, Prato: Edizioni Cassa di Risparmi e Depositi III.

Foulkes, P., Docherty, G. (2001). Variation and change in British English. Inn Hans Van de Velde & Roeland van Hout (a cura di) *'r-atics: sociolinguistic, phonetic and phonological characteristics of /r.* Brussels: ILVP. 27–43.

Giannelli L. (1976). *Toscana*. Pisa: Pacini.

Giannelli, L. (1983). Considerazioni sullo stato del rotacismo di l preconsonantico nell'Italia centrale. In *Quaderni dell'Istituto di Linguistica dell'Università di Urbino,* 1: 135–154.

Hall, T.A. (1997). *The phonology of coronals*. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.149

Ham, W. (2001). *Phonetic and phonological aspects of geminate timing*. Routledge.

Kohler, K. (1977). The production of plosives. In *Arbeitsberichtedes Instituts für Phonetik der Universität Kiel,* 8. 30-110.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in linear mixed effects models. In *Journal of Statistical Software*, 82, 13, 1-26.

Labov, W. (1966). *The Social Stratification of English in New York City* (2nd ed.). Cambridge: Cambridge University Press.

Labov, W. (1972). *Sociolinguistic patterns* (No. 4). Blackwell, Oxford.

Ladefoged, P., Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell, Oxford.

Lindau, M. (1985). The story of /r/. In Fromkin, V. (a cura di) *Phonetic linguistics: essays in honor of Peter Ladefoged*. New York: Academic Press. 157-67.

Loporcaro, M. (2009). *Profilo linguistico dei dialetti italiani*, Laterza, Bari.

Maddieson, I. (1984). The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. In *Journal of Phonetics*, 12(1), 9-15.

Mairano, P., De Iacovo, V. (2019). Gemination in northern versus central and southern varieties of Italian: A corpus-based investigation. In *Language and Speech*, 63, 3, 608-634.

Meluzzi, C., Celata, C. (2020). Style and addressees in the selection of dialectal vs. standard phonetic forms: Sicilian rhotics. In *Quaderns d'italià*, 25, 137-157.

Montemagni, S. (2008). The space of Tuscan dialectal variation: A correlation study. In *International Journal of Humanities and Arts Computing*, 2(1-2), 135-152.

Nodari, R., Meluzzi, C. (2020). Rhotic degemination in Rome Italian. In *Linguistic Studies and Essays*, 58(2), 65-98.

Nodari, R., & Calamai, S. (2021). Degemination in marginal Tuscan speech: temporal analysis in legacy speech data. In Recasens, D. & Sánchez-Miret F. (a cura di) *Sound change in Romance: phonetic and phonological issues.* Muenchen: Lincom. 67-85.

R Core Team (2021). R: A language and environment for statistical computing. [software] Versione 2024.04.2+764 https://www.R-project.org/

Romano, A. (2013). A preliminary contribution to the study of phonetic variation of /r/ in Italian and Italo-Romance. In Spreafico, L. & Vietti, A. (a cura di) *Rhotics: New Data and Perspectives*. Bolzano University Press, Bolzano, 209-225.

Romano, A., Rivoira, M., Cugno, F., Ronco, G., De Iacovo, V., & Colonna, V. (2018). Atlanti linguistici e archivi vocali di lingue locali e minoritarie in Italia. In *Géolinguistique*, (18) 1-15.

Scobbie, J. (2006). (R) as a variable. In Brown, K. (a cura di) *Encyclopaedia of Language and Linguistics (2nd ed.)*. Elsevier, Oxford, 337-344.

Solé, M.J. (2002). Aerodynamic characteristics of trills and phonological patterning. In *Journal of Phonetics* 30, 655-688.

Spinelli, A. (1981). Le comunità contadine del pratese nella lotta di liberazione e nell'assistenza ai prigionieri britannici evasi 1943-45. Fonti orali e ricerca storica nell'indagine su una classe subalterna. In *Argomenti storici*, 8, 1–27.

Spinelli, A. (1988). Archivio sonoro delle comunità contadine dell'alta val Bisenzio. In *Rassegna degli Archivi di Stato*, 48, (1-2), 232–238.

Spreafico, L., Vietti, A. (2013). *Rhotics: New data and perpectives*. Bozen-Bolzano University Press.

Sloetjs, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008) https://archive.mpi.nl/tla/elan

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. In *Language in Society*. 1(2):179-195. doi:10.1017/S0047404500000488

Van de Velde, H., Van Hout, R. (2001). *'r-atics: sociolinguistic, phonetic and phonological characteristics of /r/*. Bruxelles: ILVP.

Vietti, Alessandro, Lorenzo Spreafico & Antonio Romano. (2010). Tempi e modi di conservazione delle /r/ italiane nei frigoriferi CLIPS. In Schmid, S. (a cura di) *La dimensione temporale del parlato*. EDK. 113-127.

Wiese, R. (2001). The Phonology of /R/. In Hall, T.A. (a cura di) *Distinctive Feature Theory*. Berlin: Mouton de Gruyter. 335-368

Wiese, R. (2010). The representation of rhotics. In Van Oostendorp, M., Ewen, C.J., Hume, E., and Malden. K.R. (a cura di) *The Blackwell Companion to Phonology*. MA: Wiley-Blackwell. 1-19.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

DOI: 10.17469/O2111AISV000018

GIUDITTA AVANO, CINZIA AVESANI, MARIO VAYRA

Meccanismi di livellamento in contesto universitario: la Gorgia Toscana in soggetti fiorentini fuorisede

Levelling mechanisms in a university environment: the Gorgia Toscana in Florentine subjects studying in Bologna

The study observes the spirantization of intervocalic stops in Florentine subjects who have studied in Bologna compared to their Florentine colleagues who have studied in their hometown. The aim is to observe if moving for studying reasons made Florentine speakers' level their most known trait: the Gorgia Toscana (GT). The research investigates which linguistic and sociolinguistic factors influence levelling in a university context, considering gender, social networks, phoneme involved, lexical stress and prosodic accent. To understand if levelling in a university environment is the same as in a work migration context, we compared the results with a previous study on Florentine workers living in Bologna. The results show that living in a different town made our Florentine subjects produce lower percentages of spirantized allophones, thus it seems that they leveled GT. Levelling seems to be an unconscious process as salience and other linguistic factors don't show any effect. Sociolinguistic factors, in contrast, seem to strongly affect levelling which can be different in a university or a working migration environment.

Keywords: accommodation theory, levelling, Gorgia Toscana, sociophonetic, language contact.

1. *Introduzione*

Il presente studio propone un'analisi sociofonetica della pronuncia delle occlusive sorde intervocaliche in studenti fiorentini residenti a Bologna da cinque anni, confrontate con quella delle stesse occlusive in studenti fiorentini rimasti nella città di origine. L'obiettivo è indagare se un contatto tra varietà della stessa lingua, dovuto a migrazione per motivi di studio, porti al livellamento, ovvero alla riduzione in frequenza (Kerswill, 2013), di un tratto tipico dell'italiano fiorentino: la Gorgia Toscana (GT, Giannelli, 2000). Per comprendere quali fattori sociolinguistici e linguistici potrebbero avere un effetto sul livellamento e sul tratto stesso nei due contesti cittadini, abbiamo osservato se, oltre alla città di studio, potessero influire sulla GT, come variabili sociolinguistiche, il genere dei parlanti (Nese e Meluzzi, 2017), le reti sociali (Vietti, 2017) e l'identità di luogo (Nycz, 2018) e, come variabili linguistiche, il luogo di articolazione dell'occlusiva (Nycz, 2016), l'accento lessicale e il contesto prosodico (Foreman, 2003). Il confronto con i risultati di un nostro precedente studio (Avano, Avesani, Vayra,

in stampa), svolto con la stessa metodologia su lavoratori residenti a Bologna da più di venti anni, ci ha permesso infine di osservare se tra lavoratori e studenti fuorisede l'eventuale livellamento sia regolato dagli stessi fattori. Scopo ultimo di questo studio esplorativo è quello di valutare il contributo offerto dagli studi sul contatto dovuto a migrazione tra varietà di italiano nei confronti della ricerca sia sul contatto linguistico sia sulle varietà regionali di italiano. Il lavoro è strutturato come segue: il § 2 e il § 3 delineano rispettivamente il quadro degli studi sul contatto tra varietà di una stessa lingua dovuto a migrazione e il contatto in contesto universitario; il § 4 propone lo stato dell'arte sulla Gorgia Toscana, considerata dal punto di vista acustico e sociolinguistico. Il § 5 espone gli obiettivi della ricerca, la cui metodologia è illustrata nel § 6. Il § 7 è dedicato alla descrizione dei risultati emersi e il § 8 al confronto con i risultati del precedente studio. Seguono infine una discussione finale (§ 9), con brevi conclusioni e indicazioni sui limiti della ricerca e gli sviluppi futuri (§ 10).

2. *Migrazione, accomodamento, livellamento*

Quando un parlante si trasferisce in un'area in cui si parla un'altra lingua, ha necessariamente bisogno di apprendere il nuovo codice per comunicare; quando un parlante si sposta invece in un'area in cui si parla la stessa lingua, seppur di un'altra varietà diatopica, potrà continuare a esprimersi tramite il proprio codice ed essere compreso. Nonostante ciò, è esperienza comune osservare che, in soggetti che sono migrati in aree in cui si parla un'altra varietà della stessa lingua, il contatto tra le due varietà fa sì che la lingua del soggetto migrante tenda a modificarsi, allontanandosi dalla varietà di origine e acquisendo tratti della varietà di arrivo. Trudgill (1986) è il primo a interrogarsi sulle ragioni del fenomeno e spiega tale mutamento riprendendo la teoria sull'*accomodamento* di Giles e Howard (1973). Secondo tale teoria, si definisce accomodamento quel processo per cui i soggetti, nelle interazioni sociali, tendono, se vogliono mostrare vicinanza all'interlocutore, ad adottarne alcuni tratti. Secondo Trudgill, quando un soggetto si sposta in un'area in cui si parla una varietà differente, la continua interazione con gli abitanti del luogo porterebbe a un "accomodamento a lungo termine", dunque a un cambiamento nella competenza del parlante. Sebbene l'accomodamento "may consist of either the adoption of the new feature and/or the abandonment of the older one(s)" (Auer e Hinskens, 2005: 335) gli studi sull'accomodamento in contesto migratorio, sviluppati soprattutto nel mondo anglosassone e raccolti da Siegel (2010) sotto il nome di *Second Dialect Acquisition studies* (SDA), si sono concentrati soprattutto sull'adozione di nuovi tratti fonetici o fonologici. Tra i pochi studi sulla riduzione di tratti fonetico/fonologici della varietà di partenza, ovvero sul fenomeno del livellamento (Kerswill 2013), possiamo citare l'importante studio di Auer, Barden & Grosskopf (1998) su alcuni tratti della varietà sassone in soggetti trasferitisi nel nord della Germania. Lo scopo dello studio era indagare se (come sosteneva Trudgill, 1986) nel contatto tra varietà

l'accomodamento avvenisse solo per i tratti "salienti", ovvero i tratti di cui i soggetti sono consapevoli o se fosse invece un processo meccanico e inconsapevole che coinvolge anche i tratti non salienti. Dallo studio emerge, da una parte, come i tratti non salienti non vadano incontro a livellamento, dall'altra, come non tutti i tratti salienti vadano incontro al fenomeno. I tratti che si riducevano maggiormente in frequenza erano quelli sia salienti che stigmatizzati. In un altro studio più recente su canadesi trasferitisi a New York, (Nycz, 2016) è emerso come anche tratti non salienti vadano incontro ad accomodamento suggerendo così un processo di tipo meccanicistico. Oltre alla salienza, i fattori maggiormente indagati negli studi sulla SDA, con effetti non sempre chiari, sono l'età di arrivo (*Age of Arrival*), la durata degli anni trascorsi nel luogo di arrivo (*Length of Residence*), il genere, le reti sociali, l'attitudine e il contesto migratorio. Per una discussione sui risultati emersi si rimanda a Nycz (2015), la quale passa in rassegna gli ultimi studi sulla *Second Dialect Acquisition* con l'obiettivo di mostrare l'importanza di questo ambito di indagine da un punto di vista sociofonetico. Secondo l'autrice, gli *SDA studies* possono fornire informazioni rilevanti sulla natura e le modalità dell'apprendimento linguistico successivo all'acquisizione della prima lingua e sulla possibile evoluzione di rappresentazioni, processi e connessioni socio-indessicali nel corso della vita. Nycz sottolinea però anche come tali studi abbiano dei limiti strutturali. Il primo è legato alla difficoltà di trovare gruppi omogenei di partecipanti che abbiano le stesse caratteristiche (stessa provenienza e città di arrivo, età simili o anni di residenza simili) di modo da poter indagare i fattori che influiscono sul fenomeno senza interferenze. Il secondo è legato alla difficoltà di costruire studi longitudinali che misurino il mutamento in tempo reale, con la conseguenza che spesso si effettuano studi con intervalli di tempo troppo brevi tra una rilevazione e l'altra. Il terzo, dovuto a un tentativo di aggirare il secondo problema, è che spesso, invece di fare uno studio longitudinale, si confrontano i soggetti mobili con un gruppo di controllo che può essere composto da soggetti parlanti la varietà di arrivo o di partenza. In questi casi non si può affermare con certezza che una differenza tra chi migra e chi rimane stabile sia dovuta a livellamento: non sappiamo infatti quale fosse la varietà parlata dai soggetti prima di emigrare ed è dunque possibile che già *ab origine* non presentassero alcuni tratti della propria varietà.

3. *Varietà in contatto in contesto universitario*

Tra gli studi sull'accomodamento dovuto a migrazione, gli studi di Bigham (2010), Evans and Iverson (2007) e Pardo, Gibbons, Suppes & Krauss (2012) si sono focalizzati sul contatto tra varietà all'interno delle università, per osservare come il tipo di migrazione influisca sul fenomeno. Tra questi, lo studio di Bigham (2010) osserva come la migrazione per ragioni di studio faccia sì che il contesto del contatto sia differente da quello legato alla migrazione di tipo lavorativo, per due motivi: a) le università sono comunità intrinsecamente

temporanee; b) i soggetti provengono da più aree in cui si parlano diverse varietà della stessa lingua e tendono a vivere in una "bolla" (ivi:197), caratterizzata da uno scarso contatto con la comunità della città di arrivo e dall'essenza di una varietà dominante. In questi casi, in cui non c'è una varietà dominante verso cui convergere, i soggetti ridurrebbero i tratti più marcati della propria varietà e l'accomodamento avverrebbe nella direzione della varietà di un parlante ideale (Auer e Hinskens, 2005) meno diatopicamente marcata. La temporaneità tipica della migrazione universitaria fa però sì che non si possa parlare di *koneizzazione*: per la formazione di una vera e propria lingua di *koinè*, servirebbero infatti tre generazioni di residenza nel nuovo contesto (Kerswill, 2013). Ad oggi, l'unico studio italiano pubblicato sull'accomodamento in contesto migratorio (Nese e Meluzzi 2017) si focalizza proprio sulla peculiarità dell'accomodamento in contesto universitario. L'adozione di una pronuncia settentrionale delle affricate da parte di studenti fuorisede residenti a Pavia da sei mesi è confrontata con quella di lavoratori emigrati nella città di Bolzano, con lo scopo di verificare se processi migratori diversi per tipo, durata e prospettiva portino o meno a "un identico processo di accomodamento linguistico" (ivi, 68). Poiché in entrambi i contesti i soggetti avevano parzialmente aumentato le realizzazioni di tipo settentrionale delle affricate ed erano i soggetti femminili a guidare il mutamento, gli autori concludono che si sia di fronte a un processo di accomodamento simile che tende "verso una varietà sentita come maggiormente prestigiosa" (ivi:80). Occorre però notare che l'accomodamento nei due contesti migratori è stato analizzato con due modalità differenti (diacronia reale a Pavia *vs.* diacronia apparente a Bolzano), riguarda contesti diversi sia di partenza (vari) che di arrivo (Pavia e Bolzano) e usa corpora diversi. Lo studio sul contesto migratorio universitario, inoltre, prende in considerazione una durata di residenza nella città pavese molto breve (sei mesi). Non vi sono dunque studi nel contesto italiano che osservino il livellamento di tratti delle varietà di origine tra gli studenti fuorisede, né che osservino cosa avviene dopo un lungo periodo di permanenza. Similmente mancano studi che confrontino l'accomodamento in contesto lavorativo con quello universitario seguendo la stessa metodologia e considerando la stessa varietà di partenza. Con Nycz (2015), riteniamo che studi in tale direzione sarebbero utili per: a) osservare il mutamento dovuto ad accomodamento ad un livello fine di analisi del comportamento linguistico individuale in contesto italiano, per comprendere quali fattori lo guidino; b) osservare quali siano i rapporti tra le varietà di italiano, e come queste mutino entrando in contatto tra loro; c) comprendere e conoscere meglio linguisticamente e sociolinguisticamente anche i tratti coinvolti nel processo. Lo scopo generale del nostro studio è mostrare come gli *SDA studies* possano offrire un contributo nelle tre direzioni indicate sopra. Lo scopo specifico della ricerca, di carattere esplorativo, è esaminare come la lingua degli studenti universitari sia mutata a conclusione dell'intero ciclo di studi universitari condotti in una città diversa da quella di origine, avendo come finestra di osservazione il possibile livellamento di un tratto linguistico in un gruppo di

soggetti che condividono la stessa varietà di partenza e si sono trasferiti nella stessa città universitaria. Diversamente da Nese e Meluzzi (2017), non analizzeremo se negli studenti fuorisede l'accomodamento porti all'adozione di nuovi tratti, ma se nei fuorisede si osservi il livellamento di un tratto della varietà di origine. La varietà di origine che si è scelto di prendere in considerazione è quella fiorentina e il tratto oggetto dello studio la Gorgia Toscana. La città di arrivo è Bologna, città universitaria che ogni anno accoglie studenti fuorisede da tutta Italia (quasi il 50 % dei suoi iscritti proviene da fuori regione, fonte: Università di Bologna, http://inumeridibolognametropolitana.it/dati-statistici/istruzione/universita)[1]. Prima di esporre la metodologia del nostro studio, il paragrafo seguente è dedicato al tratto in oggetto.

4. *Gorgia Toscana*

La Gorgia Toscana (d'ora in avanti GT) è tra i tratti più studiati dell'italiano regionale toscano (Giannelli, 2000; Calamai, 2017). Fenomeno di cui si hanno tracce a partire dal XVI sec., consiste nell'indebolimento tramite spirantizzazione delle occlusive, primariamente sorde, ma anche sonore, quando si trovano in contesto postvocalico seguite da altra vocale (posizione intervocalica), da glide o da liquida (es. [la:hrima], *lacrima*). Esso porta non solo alla realizzazione di allofoni fricativi ma anche di approssimanti, giungendo fino al dileguo dell'occlusiva. Il fenomeno supera i confini di parola (es. [la hasa], *la casa*), ma sembra essere bloccato dai confini di enunciato. Di seguito alcuni esempi tipici.

(1) fuoco /ˈfuoko/ [ˈfwɔ:xo] [ˈfwɔ:ho]
foto /ˈfɔto/ [fɔ:θo]
capo /kapo/ [ka:ɸo]

Se i primi studi sulla GT si sono focalizzati sull'origine e sulla definizione del tratto (Merlo, 1950; Castellani, 1961), studi di tipo sociolinguistico hanno iniziato a svilupparsi a partire dagli anni Ottanta. Concentrandosi sull'area fiorentina, considerata il centro irradiatore del fenomeno, Giannelli e Savoia (1978) osservano come la spirantizzazione delle occlusive porti a diversi esiti collocabili su un *continuum* di graduale indebolimento "che spazia dai foni continui più aperti a quelli più chiusi e tesi" con il prevalere di foni meno aperti via via che ci si allontana da Firenze (*ivi*, 27-28). A influire sul livello di indebolimento nella realizzazione è principalmente il grado di formalità della situazione, in secondo luogo la classe sociale. Dopo questo importante studio, non sono seguite altre indagini sociolinguistiche sulla GT a Firenze. Il fenomeno

[1] La scelta della città di Bologna ci sembra particolarmente adatta, poiché le dimensioni dell'università rispetto alla città (una persona su tre è studente) e la forte presenza, tra gli studenti, di fuorisede potrebbero far sì che il contesto universitario bolognese si avvicini quello di "bolla" dei college universitari americani, con scarso contatto con la varietà di arrivo. Siamo comunque consapevoli che mediamente permane una maggiore relazione tra università e il tessuto sociale urbano rispetto al contesto universitario americano.

è invece stato studiato nella Toscana marginale, per comprendere se fosse un tratto in espansione (Cravens e Giannelli 1995; Cravens 2000). Nel loro studio su Bibbiena, Cravens e Giannelli (1995) osservano come il tratto fosse diffuso soprattutto tra i giovani e tra i soggetti maschili *"white collars"*, ipotizzando che dunque fosse un tratto caratterizzato da *covert prestige*. Secondo Cravens (poi ripreso da Villafaña Dalcher, 2008 e Marotta, 2008) la spirantizzazione della velare sarebbe l'unica saliente, stereotipo (Labov, 2010) di toscanità e più specificamente di "fiorentinità"[2]. Studi più recenti, svolti soprattutto da Calamai e colleghi tramite la tecnica del *matched guise experiment* (Biliotti e Calamai, 2010; Calamai, 2011) indicano come la varietà fiorentina in Toscana goda non più di *covert* ma di *overt prestige*, essendo associata a caratteristiche legate alla raffinatezza e all'eleganza. Preso atto che quella fiorentina in Toscana sia varietà di *overt prestige* è dunque difficile pensare che questo non valga anche per la GT, il fenomeno che maggiormente la caratterizza. Però, a causa della mancanza di uno studio di tipo sociolinguistico a Firenze, non possiamo affermare con certezza che la GT goda di tal prestigio anche nel capoluogo. Secondo Marotta (2014), che discute i risultati dell'indagine sociolinguistica di Galli de' Paratesi (1984), la GT, così come la varietà fiorentina in generale, fuori dalla regione Toscana, starebbe perdendo prestigio e sarebbe stigmatizzata specialmente da parte dei parlanti settentrionali. L'individuazione di un continuum allofonico di indebolimento delle occlusive in Giannelli e Savoia (1978) ha inaugurato un filone di studi di tipo acustico sulla GT, con indagini svolte sulle occlusive intervocaliche a Pisa (Marotta, 2001) e a Firenze (Sorianello 2003; Villafaña Dalcher, 2008). Sorianello (2003) analizza la GT nel parlato spontaneo fiorentino considerando le caratteristiche spettrografiche delle consonanti, specificamente: la presenza/assenza di barra di sonorità, burst e VOT, il rumore fricativo e il tipo di struttura formantica. In base ai risultati di questa analisi, Sorianello classifica gli esiti delle occlusive intervocaliche come occlusive, occlusive leni, occlusive sonore, fricative, approssimanti e dileguo. Marotta (2001; 2008) invece osserva come a Pisa, nel parlato letto di studenti universitari, siano presenti soltanto esiti occlusivi o fricativi, a cui aggiunge un esito intermedio, le "semifricative", allofoni caratterizzati dalla presenza di silenzio seguito da un VOT molto lungo non preceduto da scoppio (ovvero da rumore di frizione che dura più della metà del fono). Secondo l'autrice, le semifricative, intese come fricative incomplete, sarebbero presenti soltanto a Pisa perché, essendo periferica, vedrebbe una minor incidenza della spirantizzazione rispetto al capoluogo toscano, epicentro del fenomeno. Villafaña Dalcher (2008) però, nella sua indagine sul parlato letto a Firenze, osserva e inserisce tale allofono tra gli esiti fiorentini delle occlusive sorde intervocaliche. Tutti e tre gli studi confermano l'osservazione di Giannelli e Savoia (1978) secondo cui il fenomeno colpisce soprattutto il luogo di articolazione

[2] "Even among speakers who use all three spirant forms, it is only [h] which is consistently perceived consciously, while [ɸ] and [θ] normally remain below the threshold of conscious awareness. The result is that spirantization of /k/ is a Tuscan stereotype" (Cravens, 2000:14).

velare. Non vi è invece accordo rispetto al fatto se, dopo il luogo di articolazione velare, sia quello dentale (Marotta, 2001; Sorianello, 2003) o quello bilabiale (Villafaña Dalcher, 2008) a presentare maggiore spirantizzazione. Ancora meno chiara è l'influenza dei fattori contestuali: se secondo Marotta (2001) e Villafaña Dalcher (2008) l'accento lessicale non sembra influire sul grado di indebolimento delle occlusive, Sorianello rileva una maggior spirantizzazione in sillaba atona (Sorianello, Bertinetto & Agonigi, 2005). Piccardi, Ardolino (2021), similmente, osservano che le realizzazioni occlusive sono favorite in sillaba tonica, ma soltanto per il luogo di articolazione dentale. Lo studio di Sorianello, et al. (2005) è anche l'unico ad aver evidenziato un effetto del contesto prosodico frasale. Viene osservata, per le occlusive atone, "una scarsissima presenza di varianti forti se l'occlusiva è simultaneamente collocata in una parola situata alla fine di un'unità tonale" (ivi, 352) fornendo ragioni di tipo articolatorio:

> La fine di un sintagma intonativo, di natura continuativa o conclusiva, implica un naturale scadimento della pressione e dell'intensità sonora, È quindi ragionevole supporre che questa posizione prosodica, etichettabile come intrinsecamente debole, predisponga più di altre alla realizzazione di segmenti leniti" (*ibidem*).

Riassumendo, la Gorgia Toscana (GT) è un tratto presente nell'italiano regionale toscano che ha il suo centro di espansione in Firenze. È definito come un fenomeno di indebolimento tramite spirantizzazione delle occlusive, sorde e in minor grado anche sonore, in contesto post-vocalico e porta ad esiti allofonici disposti lungo un continuum. Colpisce primariamente l'occlusiva velare, mentre non è chiaro se sia influenzato dal contesto prosodico e dall'accento lessicale. È un tratto di cui i toscani sembrerebbero essere consapevoli solo per l'occlusiva velare, tanto che è noto come "c aspirata", e che sembrerebbe godere di un prestigio controverso: alto in Toscana, più basso fuori regione.

5. *Obiettivi e domande di ricerca*

Obiettivo del nostro studio è indagare se nel parlato regionale fiorentino di universitari fuorisede che hanno completato il loro percorso di studi a Bologna vi sia livellamento della GT e se questo sia regolato da fattori sociolinguistici e linguistici (RQ1). Per fare ciò, confronteremo le realizzazioni delle occlusive sorde intervocaliche di un gruppo di studenti fiorentini che vivono a Bologna (fuorisede) con quelle di un gruppo di studenti fiorentini rimasti a studiare nella città di origine. Se la GT andasse incontro a livellamento, ci aspetteremmo percentuali minori di allofoni spirantizzati nel gruppo fuorisede, che avrà invece aumentato le realizzazioni pienamente occlusive, meno diatopicamente marcate.

I fattori sociolinguistici che si è scelto di prendere in considerazione, oltre al luogo di residenza, sono a) il genere, osservato anche da Nese e Meluzzi (2017), b) le reti sociali (Milroy, 1987), fattore emerso come rilevante in alcuni tra gli SDA studies (Evans, 2004); c) "l'identità di luogo", che nello studio di Nycz

(2018) è definita come il riconoscimento e l'identificazione di un soggetto con un luogo a cui si attribuisce un significato sociale. Secondo l'autrice questo fattore avrebbe un ruolo sull'accomodamento: identificarsi maggiormente nella città di arrivo favorirebbe il processo, un forte attaccamento alla città di origine lo frenerebbe. Abbiamo considerato inoltre fattori linguistici quali il luogo di articolazione, l'accento lessicale e la posizione della parola nei costituenti prosodici. Il luogo di articolazione della consonante è una variabile considerata rilevante poiché legata al fatto che la GT sembra essere saliente, ovvero al di sopra del livello di consapevolezza dei parlanti, solo per il luogo di articolazione velare (Cravens, 2000; Marotta, 2008; Villafaña Dalcher, 2008). Vogliamo quindi verificare se solo la spirantizzazione di /k/ subisca livellamento o anche quella non saliente di /t/ e /p/(ovvero quella di cui i parlanti non sembrerebbero essere consapevoli). Sarà inoltre interessante osservare se la spirantizzazione è effettivamente maggiore per il luogo di articolazione dentale rispetto al bilabiale (cfr. § 4). L'accento lessicale e la posizione prosodica sono fattori ugualmente da considerare poiché è stato suggerito che l'accomodamento sia favorito in contesti prosodicamente prominenti, ovvero in sillaba tonica e in posizione di accento frasale (Foreman, 2003: 174). Dato che anche nella letteratura sulla GT, gli studi che hanno osservato l'influenza di tali fattori prosodici sul tratto non sono concordi (cfr. § 4), abbiamo ritenuto che includerli nel nostro studio potesse contribuire anche ad approfondire la conoscenza linguistica della GT. I risultati sono stati analizzati statisticamente con un modello generalizzato misto e un modello di regressione lineare mista per osservare quali fattori influiscano in maniera significativa sull'applicazione della Gorgia e sul grado di indebolimento delle occlusive dovuto a GT. La nostra seconda domanda di ricerca riguarda la differenza tra il livellamento osservato in contesto universitario e quello osservato nel contesto di una migrazione lavorativa a lungo termine (RQ2). A questo fine confronteremo i dati ottenuti in questo studio con quelli di un nostro studio precedente sulla Gorgia Toscana in lavoratori fiorentini (di età compresa tra i 45 e i 60 anni) residenti a Bologna da più di venti anni (Avano, *et al.*, *in stampa*). In questo caso il confronto avviene prendendo in considerazione la stessa varietà di provenienza e di arrivo. I due studi seguono inoltre la stessa metodologia e sono basati sullo stesso corpus (cfr. § 6). I lavoratori fuorisede coinvolti sono due uomini e due donne, come nel presente studio, uno laureato e uno non laureato per ogni genere. Le loro realizzazioni sono state confrontate con quelle di quattro soggetti del gruppo di controllo con le stesse caratteristiche sociolinguistiche. Dal confronto tra fuorisede e controllo era emersa una forte differenza di spirantizzazione tra chi viveva a Bologna e chi viveva a Firenze; i lavoratori residenti a Bologna tendevano a realizzare meno allofoni spirantizzati, avevano dunque, probabilmente, livellato il tratto della GT. Su tale livellamento, il genere sembrava mostrare una forte influenza: una differenza nel grado di realizzazione della GT tra lavoratori fuorisede e fiorentini sembrava essere rilevante solo per le donne. Gli uomini del gruppo di controllo e fuorisede presentavano invece

percentuali simili di spirantizzazione: non sembravano aver livellato il tratto. Anche il livello d'istruzione sembrava influire sul fenomeno: i soggetti laureati mostravano molto più livellamento dei diplomati. A Bologna, dunque, gli uomini e i soggetti diplomati tendevano verso una maggiore spirantizzazione rispetto alle donne e ai soggetti laureati. A Firenze invece, uomini e donne mostravano percentuali simili e i laureati spirantizzavano più dei diplomati, confermando il prestigio controverso della GT ipotizzato dalla letteratura (Marotta, 2014). I due soggetti lavoratori fuorisede che avevano indicato di identificarsi meno con Firenze erano quelli che presentavano meno spirantizzazione. Mentre non era stato possibile fare osservazioni legate alla rete sociale in quanto tutti e quattro i soggetti fuorisede avevano ormai legami molto allentati con la comunità di origine. I fattori linguistici non avevano mostrato invece alcun effetto sul livellamento. La riduzione del numero di allofoni spirantizzati nel gruppo fuorisede era presente per tutte e tre le consonanti coinvolte. È stata invece rilevata un'influenza del contesto prosodico sulla GT,: vi era una maggior spirantizzazione quando l'occlusiva si trovava in sillaba atona e in una parola situata in posizione finale di sintagma fonologico. Se si suppone, come in Nese e Meluzzi (2017), che la migrazione lavorativa e la migrazione universitaria portino a identici processi di livellamento, possiamo fare l'ipotesi che il livellamento che potrà emergere nei dati di questo studio sia regolato dagli stessi fattori emersi come significativi nello studio precedente.[3]

6. *Metodo*

I soggetti che abbiamo coinvolto nel presente studio sono otto. Quattro soggetti fiorentini (2M, 2F) che hanno appena concluso l'università a Bologna (e vi risiedono ancora) di età compresa tra i 24 e i 30 anni (i "fuorisede"), e quattro studenti universitari fiorentini (2M, 2F), che stanno invece compiendo i propri studi a Firenze (il "gruppo di controllo"), simili per *background* ai quattro soggetti fuorisede (Tab.1). I soggetti hanno partecipato a un esperimento di riconoscimento di immagini e di lettura di frasi. Prima è stato sottoposto loro un questionario volto ad indagare il loro *background* personale (età, età di arrivo a Bologna (*AoA*), le reti sociali, e l'identità di luogo).

[3] Nel presente studio non è però stata osservata l'influenza del livello di istruzione, essendo tutti i soggetti neolaureati. Abbiamo invece osservato l'influenza delle reti sociali, fattore non considerato nel precedente studio poiché nessuno dei soggetti aveva mantenuto contatti stretti con altri fiorentini.

Tabella 1- *Partecipanti all'esperimento, con indicazione delle risposte fornite nel questionario*

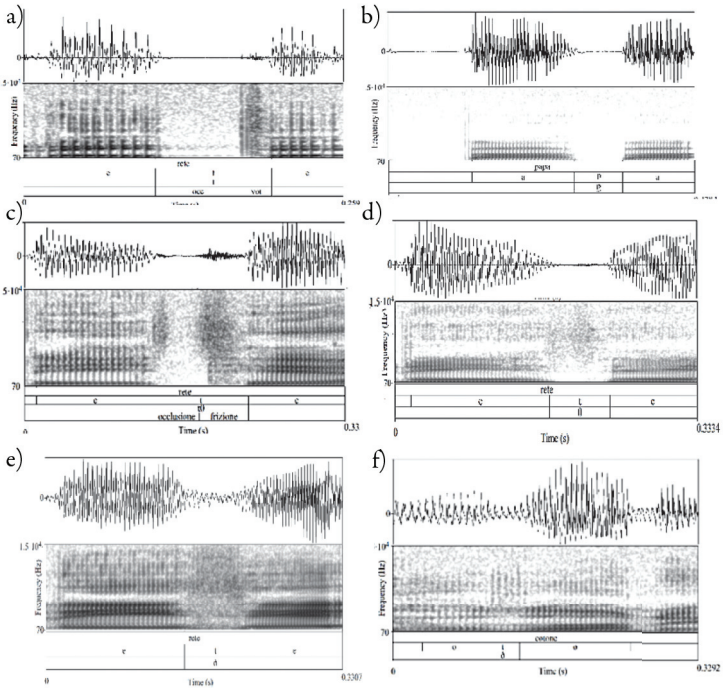
ID_ soggetto	*Età*	*AoA*	*Id Fior*	*Id Bolo*	*Vivere a Firenze*	*Contatti stretti fiorentini*	*Freq. chiamate*	*Freq. ritorno*
f_F1	24	19	3	4	no	partner	Quotidiane	1/due mesi
f_F2	29	23	3	3	no	no	Quotidiane	1/mese
f_M1	29	24	4	3	sì	Fratello, colleghi	3,4/settimana	1/settimana
f_M2	25	20	4	4	sì	no	1/settimana	1/mese
c_F1	24	/	5	/	sì	/	/	/
c_F2	20	/	4	/	sì	/	/	/
c_M1	24	/	5	/	no	/	/	/
c_M2	23	/	5	/	sì	/	/	/

Per indagare sulle reti sociali dei soggetti, è stato loro chiesto di indicare la provenienza dei dieci contatti a loro più vicini, per vedere quanti tra questi fossero di Firenze, la frequenza delle telefonate con persone rimaste a Firenze ("Freq. chiamate") e la frequenza di ritorno a casa ("Freq. ritorno"). Per indagare l'influenza dell'identità di luogo sulla spirantizzazione abbiamo chiesto di indicare quanto (da 1 a 5) si identificassero sia con Bologna ("Id Bolo") sia con Firenze ("Id Fior"). Ai soggetti fuorisede si chiedeva se volessero tornare a Firenze, ai soggetti del gruppo di controllo se volessero continuare a vivere in città. Il corpus per il compito di lettura e riconoscimento di immagini consisteva in 60 parole piane, tutte sostantivi, contenenti occlusive sorde intervocaliche, scelti dal Nuovo Vocabolario di Base di De Mauro (2016), quindi tra le 6000 parole più frequenti dell'italiano. Vi erano venti parole per ogni luogo di articolazione, di cui dieci in cui l'occlusiva si trovava in sillaba atona finale (es. <foto>) e dieci in cui si trovava in sillaba tonica prefinale (es. <cotone>). Ognuna di queste parole è stata pronunciata quattro volte, la prima in isolamento nel compito di riconoscimento di immagini, le altre tre durante la lettura di frasi in cui la parola si trovava in posizione finale di un diverso costituente prosodico. Le posizioni considerate, di forza prosodica crescente, sono le seguenti: finale di sintagma fonologico (ϕ), finale di sintagma intonativo (*I*) e finale di enunciato (*U*) e si rifanno alla teoria della *Prosodic phonology* (cfr. Nespor, Vogel, 2007)[4]. Di seguito riportiamo un esempio delle frasi del *task* di lettura relativo alla parola <rete>, posta nelle tre diverse posizioni finali rispetto ai costituenti prosodici:

(2) ϕ [[[La **rete**]$_{ϕ}$ è il sistema di pesca più antico]$_{I}$]$_{U}$
I [I pesci erano rimasti intrappolati [nella r**ete**]$_{ϕ}$]$_{I}$[ma si muovevano ancora]$_{I}$]$_{U}$
U [[[Marco ha trovato un buco [nella **rete**]$_{ϕ}$]$_{I}$]$_{U}$

[4] In particolare, per una revisione critica relativa al sintagma fonologico e alla rilevanza di questo costituente per l'applicazione di fenomeni segmentali nei dialetti e nelle varietà di italiano si rimanda a Vogel (1999:62-65).

Figura 1 - *Esiti allofonici dovuti a GT: a) occlusivo, b) lenito, c) semifricativo, d) fricativo sordo, e) fricativo sonoro, f) approssimante*



Il corpus analizzato consta di 1920 occorrenze (60 parole x 4 ripetizioni x 4 parlanti), di cui sette sono state scartate[5]. Le registrazioni sono state effettuate nelle abitazioni dei soggetti con un microfono ad archetto *Shure WH20* connesso al registratore da campo *Zoom handy4 recorder* con un campionamento a 48 khz e formato 24bit. L'analisi spettrografica (effettuata con *Praat,* Boersma, Weenink, 2023) ha portato alla classificazione degli esiti delle occlusive sorde intervocaliche seguendo e adattando la classificazione allofonica proposta negli studi acustici di Marotta (2001), Sorianello (2003) e Villafaña Dalcher (2008). Le occorrenze sono state classificate come:

– occlusive – definite dalla fase di silenzio seguita da un burst e un VOT;
– lenite (bilabiali) – definite dall'assenza di VOT e burst, eventualmente con tracce di sonorità durante la fase di occlusione[6];
– semifricative[7] - definite da una fase di silenzio seguita da rumore di frizione (VOT lungo più della metà del fono senza scoppio);

[5] In tre casi l'occorrenza è stata scartata per impossibilità di analisi spettrografica. In quattro casi, la consonante è andata incontro a dileguo. Dato che i quattro casi erano tutti relativi al parlante c_M3, si è deciso di considerare tale esito una variante individuale.

[6] Allofoni leniti, come in Sorianello (2003) sono stati individuati soltanto per il luogo di articolazione bilabiale.

[7] Inizialmente, seguendo Sorianello (2003) e Marotta (2008), non avevamo inserito l'allofono semifricativo nella classificazione degli esiti della GT a Firenze. Abbiamo però effettivamente rilevato casi

– fricative sorde - definite dalla presenza di rumore per tutta la durata del fono;
– fricative sonore - definite dalla presenza di rumore per tutta la durata del fono e barra di sonorità;
– approssimanti - definite dalla presenza di un chiaro andamento formantico e da una riduzione di intensità (evidenziata dall'oscillogramma) rispetto alle vocali adiacenti. Esempi di ciascuno di questi allofoni sono riportati in Fig. 1. Gli allofoni sono da considerarsi come esiti di un indebolimento crescente dei rispettivi fonemi occlusivi.

7. *Risultati*

7.1 Distribuzione degli allofoni

Nella Tab. 2 riportiamo le realizzazioni delle occlusive sorde intervocaliche distinte per gruppo[8]. Nel gruppo dei fuorisede la percentuale di occlusive sorde risulta molto più alta rispetto al gruppo di controllo (45 % vs 10 %), mentre la percentuale degli allofoni indeboliti risulta molto minore (semifricative: 5 % vs 10 %; fricative sorde 31 % vs 49 %; fricative sonore 7 % vs 19 %). Sembra dunque che i soggetti fuorisede presentino un minor grado di spirantizzazione delle occlusive sorde intervocaliche. Nelle Tab. 3 e 4 i dati precedenti sono scorporati per genere e soggetto. Il genere (Tab. 3) non sembra avere un effetto chiaro né sulla GT né sul suo livellamento: gli uomini, se da un lato realizzano più esiti forti delle donne, ovvero più occlusive, dall'altra realizzano anche più approssimanti, l'esito più indebolito. Le donne privilegiano invece gli esiti intermedi: le fricative. L'effetto è più marcato nel gruppo di controllo dove le due donne realizzano appena il 3 % di occlusive contro il 17 % dei due uomini, e il 57 % di fricative sorde contro il 36 % dei due uomini. La differenza tra soggetti fuorisede e controllo è maggiore tra le donne: le donne fuorisede realizzano il 39 % di occlusive in più e il 23 % di fricative sorde in più rispetto alle donne del gruppo di controllo (42 % vs 3 % e 34 % vs 57 %); gli uomini fuorisede realizzano il 31 % di occlusive in più e il 7 % di fricative in meno rispetto ai due uomini del gruppo di controllo.

Se osserviamo però la distribuzione allofonica nei singoli parlanti (Tab. 4), appare evidente come ì in entrambi i gruppi vi sia una variabilità individuale molto forte che il genere non pare in grado di spiegare.

---

in cui vi era una breve fase di occlusione seguita da un rumore molto lungo non preceduto da scoppio, difficilmente classificabile come VOT. Ipotizziamo che le semifricative siano presenti nell'italiano fiorentino del nostro studio e in quello di Villafaña Dalcher (2008), perché entrambi analizzano il parlato letto, ovvero un parlato più controllato rispetto a quello spontaneo analizzato da Sorianello (2003).

[8] Nelle Tab. 3, 4, 5 e 6 i dati del task di lettura e della pronuncia in isolamento sono stati accorpati, in quanto abbiamo considerato la parole pronunciate in isolamento come frasi monorematiche. Le realizzazioni allofoniche relative alla pronuncia delle parole in isolamento sono state distinte da quelle delle parole del task di lettura, situate in diverse posizioni all'interno dei costituenti prosodici, in Tabella 7.

Tabella 2 - *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte per gruppo*

	occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
fuorisede	430 (45 %)	41 (4 %)	51 (5 %)	300 (31 %)	65 (7 %)	69 (7 %)
controllo	95 (10 %)	24 (3 %)	107 (11 %)	446 (47 %)	185 (19 %)	100 (10 %)

Tabella 3 - *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte per gruppo e genere (F = donne, M = uomini)*

		occlusive	*lenite*	*semifricative*	*Fricative sorde*	*Fricative sonore*	*approssimanti*
fuorisede	F	201 (42 %)	39 (8 %)	31 (6 %)	162 (34 %)	35 (7 %)	11 (2 %)
	M	229 (48 %)	2 (0 %)	20 (4 %)	138 (29 %)	30 (6 %)	58 (12 %)
controllo	F	12 (3 %)	8 (2 %)	47 (10 %)	272 (57 %)	99 (21 %)	42 (9 %)
	M	83 (17 %)	16 (3 %)	60 (13 %)	174 (36 %)	86 (18 %)	58 (12 %)

Nel gruppo di fuorisede il 48 % di occlusive prodotte dagli uomini è dato dalla media tra il 2 % di occlusive realizzate da f_M1 e il 93 % di f_M2. F_M1 d'altronde non realizza alcun esito né fricativo né approssimante, mentre f_M2 realizza fricative e approssimanti nel 95 % dei casi. Nel gruppo delle donne fuorisede c'è moltissima differenza: f_F1 realizza soltanto l'11 % di occlusive e presenta alte percentuali di fricative e approssimanti, f_F2 realizza il 73 % di occlusive e non produce alcuna approssimante. Se è vero che in una prima fase dopo la migrazione è noto che vi sia una fortissima variabilità individuale nel livellamento, legata all'abilità dei singoli parlanti nel gestire il tratto (Kerswill, 2013:528), tra i soggetti fuorisede vi sono comunque due comportamenti: f_F1 e f_M1 realizzano basse percentuali di occlusive, alte di fricative (oltre il 50 %) e realizzano anche allofoni sonori; fricative sonore e approssimanti. f_F2 e M2 invece realizzano percentuali oltre il 70 % di occlusive, f_F2 realizza il 16 % di fricative contro lo 0 % di f_M1 ma anche f_F2 non realizza praticamente allofoni più indeboliti delle fricative sorde. Confrontando questi dati con le risposte fornite dai soggetti (Tab. 1) sui loro contatti con i fiorentini e sulla loro identità di luogo osserviamo che i due soggetti che presentano più spirantizzazione, f_M1 e f_F1, sono quelli che hanno mantenuto contatti più stretti con fiorentini. F_M1 ha indicato di vedere spesso il fratello, a Bologna anche lui, di avere colleghi toscani e di tornare a casa tutti i fine settimana. F_F1 invece, anche se torna raramente a casa, ha il partner fiorentino che vede una volta a settimana e che sente quotidianamente a telefono. Gli altri due soggetti, che presentano percentuali molto alte di occlusive, non hanno invece indicato contatti stretti con fiorentini, non tornano spesso a casa e a telefono sentono solo i genitori (per f_M2 anche raramente). Le risposte fornite riguardo all'identità di luogo sembrano dare indicazioni meno chiare: f_F1, l'unica che si sente più bolognese che fiorentina, presenta un grado di spirantizzazione molto più forte di chi, come f_M2 o f_F2, si sente egualmente fiorentino e bolognese.

Tabella 4 - *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte prima per gruppo (f= fuorisede, c = controllo), e per genere (F = donne, M = uomini) e infine per singolo parlante*

	occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
f_F	201 (42 %)	39 (8 %)	31 (6 %)	162 (34 %)	35 (7 %)	11 (2 %)
f_F1	26 (11 %)	31 (13 %)	15 (6 %)	124 (52 %)	33 (14 %)	11 (5 %)
f_F2	175 (73 %)	8 (3 %)	16 (7 %)	38 (16 %)	2 (1 %)	0 %
f_M	229 (48 %)	2 (0 %)	20 (4 %)	138 (29 %)	30 (6 %)	58 (12 %)
f_M1	5(2 %)	2 (1 %)	6 (2 %)	138 (57 %)	30 (12 %)	58 (24 %)
f_M2	224(94 %)	0 %	14 (6 %)	0 %	0 %	0 %
c_F	12(3 %)	8 2 %)	47(10 %)	272(57 %)	99(21 %)	42(9 %)
c_F1	12(5 %)	4 (2 %)	41(17 %)	171(71 %)	8(3 %)	4(2 %)
c_F2	0 %	4(2 %)	6(3 %)	101(42 %)	91(38 %)	38(16 %)
c_M	83(17 %)	16(3 %)	60(13 %)	174(36 %)	86(18 %)	58(12 %)
c_M1	83(35 %)	16(7 %)	57(24 %)	83(35 %)	0 %	1(0 %)
c_M2	0 %	0 %	3(1 %)	91(38 %)	86(36 %)	57(24 %)

Anche nel gruppo di controllo notiamo che il genere non sembra un fattore utile per spiegare la variabilità. Osserviamo che c_F2 e c_M2 presentano un livello molto avanzato di spirantizzazione: nessuna occlusiva, quasi il 40 % fricative sonore e percentuali alte di approssimanti. C_M1 e c_F1 invece presentano una certa percentuale di occlusive, bassa per c_F1, oltre il 30 % per c_M1. Confrontando sempre i risultati con le risposte fornite dai soggetti, si potrebbe ipotizzare un effetto dell'identità di luogo. C_F1 è l'unica a non aver affermato di sentirsi pienamente fiorentina, c_M1 invece è l'unico ad avere detto di voler andare via da Firenze, mostrando poco attaccamento alla città. C_F2 e c_M2 invece avevano indicato entrambi di sentirsi massimamente fiorentini e di voler rimanere in città. Passando invece ai fattori linguistici, osservando innanzitutto le realizzazioni allofoniche, distinte tra sillaba atona e sillaba tonica (Tab. 5), si nota come il fatto che l'occlusiva intervocalica si trovi in sillaba atona o tonica non sembra influire sul suo indebolimento in nessuno dei due gruppi, dunque nemmeno sul livellamento del tratto. Come atteso, il luogo di articolazione ha un effetto sulla GT (Tab. 6): in entrambi i gruppi la velare è la consonante che spirantizza maggiormente (nel gruppo di controllo vi è appena l'1 % di occlusive), ed è l'unica a presentare alte percentuali di esiti in cui è coinvolta anche la sonorità (fricative sonore e approssimanti). La differenza tra i due gruppi nell'applicare la GT è simile per tutti e tre i luoghi di articolazione. Le occlusive nel gruppo di fuorisede, infatti, aumentano di circa 35 % rispetto al gruppo di controllo sia per /k/ sia per /t/ sia per /p/.All'interno di entrambi i gruppi, le percentuali di occlusive per le bilabiali e le dentali sono simili, così come quelle delle fricative sorde. Soltanto nel gruppo di controllo si osserva una percentuale leggermente più alta di fricative sonore per il luogo di articolazione dentale (21 % contro 13 % di /p/). Nella Tab. 7 riportiamo le realizzazioni allofoniche distinte per gruppo e posizione della parola entro i costituenti prosodici: sintagma fonologico (φ), sintagma intonativo (I) ed Enunciato (U).

Tabella 5 - *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte per gruppo e sillaba*

		occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
fuorisede	atona	212 (45 %)	21 (4 %)	21 (4 %)	152 (32 %)	41 (9 %)	29 (6 %)
	tonica	218 (45 %)	20 (4 %)	30 (6 %)	148 (31 %)	24 (5 %)	40 (8 %)
controllo	atona	45 (9 %)	11 (2 %)	51 (11 %)	239 (50 %)	75 (16 %)	58 (12 %)
	tonica	50 (10 %)	13 (3 %)	56 (12 %)	207 (43 %)	110 (23 %)	42 (9 %)

Tabella 6 - *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte per fonema e per gruppo*

		occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
/k/	fuorisede	112 (35 %)	0 %	29 (9 %)	92 (29 %)	27 (8 %)	58 (18 %)
	controllo	4 (1 %)	0 %	23 (7 %)	135 (43 %)	76 (24 %)	79 (25 %)
/t/	fuorisede	170 (53 %)	0 %	15 (5 %)	106 (33 %)	20 (6 %)	8 (3 %)
	controllo	48 (15 %)	0 %	42 (13 %)	150 (47 %)	68 (21 %)	12 (4 %)
/p/	fuorisede	148 (46 %)	41 (13 %)	7 (2 %)	102 (32 %)	18 (6 %)	3 (1 %)
	controllo	43 (13 %)	24 (8 %)	42(13 %)	161 (50 %)	41 (13 %)	9 (3 %)

Tabella 7- *Realizzazioni allofoniche delle occlusive sorde intervocaliche distinte per posizione della parola entro i costituenti prosodici e per gruppo*

		occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
fuorisede	isolamento	112 (47 %)	11 (5 %)	10 (4 %)	75 (31 %)	18 (8 %)	13 (5 %)
	ϕ	106 (44 %)	8 (2 %)	13 (5 %)	57 (24 %)	27 (11 %)	32 (13 %)
	I	110 (46 %)	5 (3 %)	13 (5 %)	84 (35 %)	11 (5 %)	13 (5 %)
	U	102 (43 %)	17 (7 %)	15 (6 %)	84 (35 %)	9 (4 %)	11 (5 %)
controllo	isolamento	26 (11 %)	2 (1 %)	31 (13 %)	120 (50 %)	41 (17 %)	22 (9 %)
	ϕ	29 (12 %)	3 (1 %)	18 (8 %)	93 (39 %)	59 (25 %)	35 (15 %)
	I	26 (9 %)	10 (4 %)	21 (9 %)	10 (65 %)	53 (22 %)	27 (11 %)
	U	19 (8 %)	9 (4 %)	37 (15 %)	127 (53 %)	32(13 %)	16 (7 %)

La percentuale di occorrenza delle occlusive, l'allofono forte, non mostra sostanziali variazioni in funzione della posizione nella struttura prosodica sia per il gruppo dei fuorisede (la percentuale varia tra il 43 e il 47 %) che per quello di controllo (la percentuale varia tra l'8 e il 12 %). Più variabili, invece, a seconda della posizione, le realizzazioni degli allofoni indeboliti. Per entrambi i gruppi di parlanti sembra di poter delineare due tendenze diverse legate alla sonorità degli allofoni: per gli allofoni sordi (lenite, semifricative, fricative sorde), la percentuale di realizzazione aumenta con l'aumentare della forza della posizione prosodica: la percentuale

maggiore di allofoni indeboliti è prodotta in posizione finale di Enunciato (U, con una sola eccezione). Per gli allofoni sonori (fricative sonore e approssimanti), la percentuale di realizzazione diminuisce con l'aumentare della forza della posizione prosodica: il numero maggiore di realizzazioni indebolite è prodotto in posizione prosodicamente debole (φ) e diminuisce in posizione finale di I e U.

7.2 Analisi statistica

Considerata la forte variabilità interpersonale, per avere conferma del fatto che i soggetti fiorentini residenti a Bologna applichino meno la Gorgia dei soggetti in sede a causa del contesto di contatto in cui hanno vissuto abbiamo applicato ai dati (N = 1913) un modello generalizzato misto con variabile dipendente nominale a due livelli: assenza (occlusive) o presenza di GT (considerando tutti gli allofoni indeboliti come casi in cui si è applicata la GT). Come variabili fisse nel modello si è inserito il Gruppo in interazione con il Genere, il Luogo di articolazione, l'Accento lessicale e la Posizione della parola nei costituenti prosodici. I parlanti sono stati considerati come fattore random. In Tab. 8 il modello[9] che meglio rappresenta i dati secondo l'*Akaike Information Criterion* (AIC). Il Gruppo (fuorisede vs controllo) sembra essere un fattore significativo nel predire la percentuale di esiti con GT (p=0.038), così come il fonema (p<0.001). Il gruppo di fuorisede realizza molti più esiti non indeboliti del gruppo di controllo anche eliminando l'influenza del singolo parlante. Similmente la GT si conferma un tratto su cui il luogo di articolazione esercita la sua influenza.

Tabella 8 - *Modello generalizzato misto con variabile dipendente a due livelli (non applicazione della GT=0/ applicazione=1). I livelli di riferimento sono il fonema /p/ e il gruppo di controllo, l'intercetta corrisponde al logit (applicazione/non applicazione) della GT per /p/ nel gruppo di controllo)*

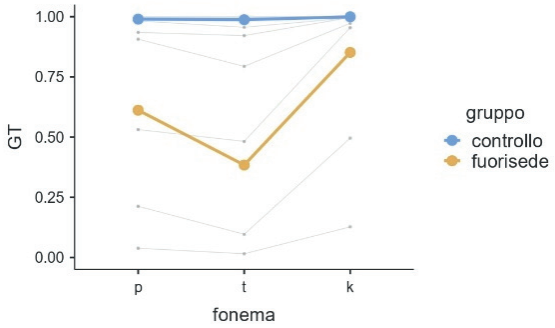
Variabile	*Estimate*	*Standard Error*	*Odd ratio*	*z*	*p*
(Intercetta)	3,024	1,183	20,56840	2,56	,011
Gruppo [fuorisede – controllo]	-4,895	2,364	0,00749	-2,07	,038
Fonema [t – p]	-0,562	0,204	0,57033	-2,75	,006
Fonema [k – p]	2,104	0,308	8,19871	6,83	< ,001
gruppo*fonema [fuorisede - controllo * t – p]	-0,730	0,408	0,48174	-1,79	,073
gruppo*fonema [fuorisede - controllo * k – p]	-1,623	0,616	0,19725	-2,63	,008

Nessuno degli altri fattori è invece emerso come significativo (non influisce sull'applicazione della Gorgia). L'unica interazione emersa come significativa è

[9] Il modello generalizzato misto, così come il modello misto seguente, fanno parte del pacchetto statistico GAMLj (Gallucci, 2019) e sono stati sviluppati sul *software* Jamovi (2021).

quella tra gruppo e fonema (p =.018). Dal grafico (Fig. 2) si nota che ciò avviene perché nel gruppo di controllo la GT è in uno stato molto avanzato per tutti e tre i luoghi di articolazione (anche la bilabiale, che spirantizza meno della velare, ha una realizzazione occlusiva nel 13 % dei casi).

Figura 2 - *Visualizzazione grafica dell'interazione tra gruppo e fonema sull'applicazione della GT*



Tra i fuorisede invece la differenza di spirantizzazione tra velare e gli altri due luoghi di articolazione rimane più ampia. La velare non spirantizza nel 35 % dei casi, contro il 54 % della dentale. Tale interazione indica, dunque, che la velare, per cui la GT è saliente, è il luogo di articolazione per cui la GT è andata incontro a minor livellamento. Per osservare l'effetto degli stessi fattori sul grado di indebolimento consonantico delle occlusive, (seguendo Bross, 2019) abbiamo poi applicato ai dati un modello di regressione lineare mista, considerando, come variabile dipendente ordinale, gli esiti allofonici. Questi ultimi sono stati interpretati come gradi di un continuum di indebolimento consonantico crescente i cui estremi sono le occlusive e le approssimanti. Per essere inseriti nel modello di regressione lineare sono stati ricodificati come valori in scala da 0 a 5 (occlusive =0, lenite =1, semifricative =2, fricative =3, fricative sonore =4, approssimanti =5). Come nel modello precedente, i parlanti sono stati considerati un fattore *random*. Il modello (Tab. 9) ha messo in luce come la enorme variabilità intersoggettiva incida sulla significatività del Gruppo, fattore che non risulta statisticamente significativo per il grado di indebolimento delle occlusive. Invece, sia il Luogo di articolazione (p<.001), che la Posizione della parola nei costituenti prosodici (p<.001) risultano fattori altamente significativi. Per quanto riguarda il Luogo di articolazione si osserva che il grado di indebolimento delle occlusive non solo è maggiore per /k/ rispetto a /p/ ma anche per /t/ rispetto a /p/ (t=2.65, p =.008). Infine, per quanto riguarda il contesto prosodico, si osserva che l'indebolimento delle occlusive è significativamente maggiore in posizione finale di sintagma fonologico rispetto a quella di sintagma intonativo (p=.002) e in posizione finale di sintagma intonativo rispetto alla posizione finale di enunciato (p=.033). L'indebolimento è dunque sensibile alla posizione della parola nel costituente prosodico, in ragione inversa rispetto alla forza del costituente

Tabella 9 - *Modello lineare misto con variabile dipendente continua, considerando gli esiti allofonici come gradi diversi di indebolimento consonantico. Osservazioni: 1913, variabile random: Parlante. L'intercetta corrisponde al grado di indebolimento delle bilabiali (/p/ in posizione finale di sintagma intonativo (I)*

Variabile	*Estimate*	*StandardError*	*t*	*p*
(Intercept)	2,3348	0,4955	4,71	,002
Posizione [*U* – *I*]	-0,1241	0,0581	-2,13	,033
Posizione [*φ* – *I*]	0,1838	0,0582	3,16	,002
Posizione [*0* – *I*]	-0,0698	0,0580	-1,20	,229
Phoneme [/t/ – /p/]	0,1334	0,0502	2,66	,008
Phoneme [/k/ – /p/]	1,0194	0,0503	20,26	< ,001

8. *Confronto con il livellamento in contesto migratorio lavorativo*

Il fatto che nel presente studio i soggetti siano tutti universitari o neolaureati mentre nel precedente (Avano *et al.*, *in corso di stampa*) i soggetti fuorisede erano metà laureati e metà diplomati, consente di fare un confronto parziale dei risultati, dopo aver distinto il gruppo di lavoratori fuorisede e di lavoratori in sede per livello d'istruzione (Tab. 10). Il confronto rivela che le percentuali degli studenti universitari fuorisede sono molto simili a quelle dei due lavoratori laureati fuorisede, nonostante i lavoratori abbiano molti più anni di residenza a Bologna. Gli studenti fuorisede presentano infatti appena il 9 % in meno di realizzazioni occlusive (45 % contro 54 %), la stessa percentuale di approssimanti, il 3 % di fricative sonore in più e il 10 % in più di fricative sorde (31 % contro 22 % dei lavoratori).

Tabella 10 - *Realizzazioni allofoniche delle occlusive intervocaliche sorde realizzate da studenti fuorisede e controllo confrontate con quelle dei lavoratori fuorisede e controllo, distinti per livello di istruzione*

		occlusive	*lenite*	*semifricative*	*fricative sorde*	*fricative sonore*	*approssimanti*
fuorisede	Studenti	430 (45 %)	41 (4 %)	**51 (5 %)**	**300 (31 %)**	**65 (7 %)**	**69 (7 %)**
	Lavoratori	**352 (36 %)**	**41 (4 %)**	**105 (11 %)**	**357 (37 %)**	**36 (4 %)**	**77 (8 %)**
	Laureati	259 (54 %)	8 (2 %)	53 (11 %)	103 (22 %)	21 (4 %)	34 (7 %)
	Diplomati	83 (17 %)	33 (7 %)	52 (11 %)	254 (53 %)	15 (3 %)	43 (9 %)
controllo	Studenti	**95(10 %)**	**24 (3 %)**	**107 (11 %)**	**446 (47 %)**	**185 (19 %)**	**100 (10 %)**
	Lavoratori	**135 (14 %)**	**70 (7 %)**	**161 (17 %)**	**505 (53 %)**	**26 (3 %)**	**62 (6 %)**
	Laureati	42 (9 %)	17 (4 %)	96 (20 %)	257 (54 %)	20 (4 %)	48 (10 %)
	Diplomati	94 (20 %)	53 (11 %)	65 (14 %)	248 (52 %)	6 (1 %)	14 (3 %)

I lavoratori diplomati fuorisede invece mostrano percentuali di spirantizzazione molto più alte dei nostri studenti, nonostante vivano fuori da Firenze da molto più tempo (appena il 17 % di occlusive, il 53 % di fricative sorde). Anche a Firenze

le percentuali degli studenti e dei lavoratori laureati sono simili, mentre sono i lavoratori diplomati a spirantizzare meno di tutti.

9. *Discussione*

Dal nostro studio è emerso che gli studenti che vivono a Bologna realizzano percentuali di allofoni spirantizzati nettamente inferiori rispetto ai loro colleghi rimasti a studiare nella sede fiorentina, e la differenza risulta statisticamente significativa. Potrebbe essersi dunque verificato un livellamento del tratto della GT, similmente a quanto avvenuto tra i lavoratori fuorisede del nostro precedente studio. A differenza di quest'ultimo, e dello studio di Nese e Meluzzi (2017) nel quale erano le donne a guidare il mutamento verso una varietà percepita come più prestigiosa, non è emerso un effetto del genere né sul livellamento né sull'applicazione o il grado di indebolimento delle occlusive dovuto a GT. La differenza di risultati rispetto allo studio di Nese e Meluzzi potrebbe essere spiegata considerando che nel caso degli studenti di Pavia le due varietà in contatto erano caratterizzate da una differenza di prestigio maggiore rispetto al nostro studio: si ipotizza infatti che il fiorentino goda di un prestigio, sebbene in calo, superiore a quello delle varietà meridionali studiate da Nese e Meluzzi. È più difficile invece spiegare la mancanza di un effetto del genere rispetto al nostro studio precedente. Si potrebbe ipotizzare che il prestigio del fiorentino sia cresciuto tra il momento in cui i lavoratori si sono trasferiti a Bologna (primi anni duemila) e quello in cui gli studenti universitari hanno fatto lo stesso (attorno al 2015). Tale ipotesi è però in contrasto con l'osservazione che il livellamento degli studenti universitari è molto simile a quello dei lavoratori laureati e maggiore di quello dei lavoratori diplomati (cfr. Tab. 10), fatto, questo, che suffragherebbe l'ipotesi che la GT non godesse e ancora non goda di *overt prestige* fuori dai confini regionali. Ci sembra più probabile un'altra spiegazione: in una varietà non stigmatizzata quale è il fiorentino inserita nel contesto universitario bolognese, il genere non risulta più un fattore affidabile per valutare il prestigio di un tratto. Donne e uomini potrebbero avvertire la stessa pressione verso una pronuncia meno marcata e più prestigiosa che tenda a non spirantizzare le occlusive sorde intervocaliche. Un fattore che meglio permette di comprendere lo status di prestigio della GT e di prevedere il suo livellamento è invece il livello di istruzione dei parlanti. È interessante notare che non vi sono studi sull'accomodamento dovuto a migrazione che tengono in considerazione il livello di istruzione dei soggetti migrati. Il fattore che meglio potrebbe spiegare la grande variabilità di realizzazioni tra gli studenti fuorisede è la loro rete sociale in linea con quanto emerso in alcuni studi sulla *Second Dialect Acquisition* (Evans, 2004): chi ha mantenuto più contatti con Firenze e con i fiorentini tende a spirantizzare di più, ovvero a livellare meno il tratto. L'effetto delle reti sociali, e in particolare del mantenimento dei contatti, non era stato indagato da Nese e Meluzzi (2017) sugli studenti di Pavia, né era stato possibile indagarne l'effetto nel nostro studio precedente perché, dopo vent'anni, nessuno aveva mantenuto più contatti stretti

con soggetti della città di origine. L'identità di luogo sembra invece poter spiegare la variabilità solo all'interno del gruppo di controllo, in contrasto con il nostro studio precedente. La risposta alla nostra seconda domanda di ricerca è quindi che i due contesti migratori sembrano aver indotto un livellamento regolato da fattori sociolinguistici differenti. Per quanto riguarda i fattori linguistici, i risultati di entrambi gli studi hanno mostrato che il livellamento è influenzato dal luogo di articolazione (/k/mostra minor livellamento), ma non dalla prominenza accentuale. Ciò permette di ipotizzare che per universitari e lavoratori il processo di accomodamento sia lo stesso e sia un processo probabilmente inconscio e meccanico (Nycz, 2016). Poiché la GT sembrerebbe essere saliente solo per il luogo di articolazione velare, se andassero incontro a livellamento solo i tratti salienti avremmo dovuto osservare, al contrario, una riduzione della spirantizzazione nei gruppi fuorisede solo per /k/. Il livellamento ha invece riguardato la spirantizzazione di tutte e tre le occlusive, addirittura in maniera minore per la velare rispetto agli altri due luoghi di articolazione. Relativamente all'influenza dei fattori linguistici sulla GT, notiamo innanzitutto come il luogo di articolazione influisca sia sulla applicazione della GT, sia sul grado di indebolimento delle occlusive. La GT si applica maggiormente a /k/, quindi a /p/ e in modo minore a /t/ in entrambi i gruppi (ma in modo più evidente nel gruppo dei fuorisede). È sempre la velare ad andare incontro a maggior indebolimento, in questo caso seguita dalla dentale e quindi dalla bilabiale. Il fatto che in letteratura non fosse chiaro se spirantizzasse maggiormente /p/ o /t/ dipende dunque da ciò che si osserva: la GT si applica in percentuale maggiore a /p/ che a /t/, ma il grado di indebolimento, una volta che la GT sia stata applicata, è maggiore per /t/ rispetto a /p/. In altre parole, /t/ presenta percentuali maggiori di fricative sonore e approssimanti rispetto a /p/. La posizione della parola nei costituenti prosodici non influisce sull'applicazione della GT, ma influisce sul grado di indebolimento delle occlusive: posizioni più forti dal punto di vista prosodico (posizione finale di sintagma intonativo e di enunciato, ma non di sintagma fonologico) sembrano limitare la realizzazione degli allofoni più deboli. Il fatto che l'accento di sintagma fonologico non sembri limitare, invece, la realizzazione di allofoni più deboli sembrerebbe in favore dell'ipotesi che in italiano regionale toscano il sintagma fonologico non sia un dominio rilevante per i fenomeni di tipo segmentale. Secondo Vogel (1999:65) in toscano anche il Raddoppiamento Fonosintattico, infatti, non sarebbe bloccato dal confine di sintagma fonologico, per cui il suo dominio sembrerebbe essere più ampio. L'accento lessicale non ha invece mostrato alcun effetto, in linea con gli altri studi sul parlato letto (Marotta 2001 e Villafaña Dalcher 2008), ma diversamente da quanto emerso negli studi sul parlato spontaneo (Sorianello *et al.*, 2005; Piccardi e Ardolino 2021) per il luogo di articolazione dentale. Occorrerebbe dunque approfondire se tale differenza sia legata ai due diversi tipi di parlato.

10. *Conclusioni, limiti e future ricerche*

Negli studenti residenti a Bologna sembrerebbe essersi verificato un livellamento della GT su cui non influisce il genere, né alcun fattore linguistico, ma su cui invece sembrerebbero influire le reti sociali (RQ1). Il cambiamento di contesto di migrazione, da lavorativo a universitario, sembrerebbe portare a un processo di livellamento identico dal punto di vista linguistico (inconscio), ma su cui i fattori sociolinguistici influiscono diversamente (RQ2). La GT, tratto bandiera di una varietà, quella fiorentina, sebbene non stigmatizzata nel panorama nazionale e avente alto prestigio all'interno dei confini regionali, sembra essere percepita come un tratto da ridurre in frequenza tra i soggetti che emigrano e principalmente da quelli con un alto livello di istruzione. Il nostro studio introduce la novità di un effetto del contesto prosodico sul tratto e chiarisce l'effetto della GT sulle occlusive bilabiali e dentali. La nostra ricerca presenta tutti i limiti messi in luce da Nycz (2015) per gli altri studi sul contatto tra varietà dovuto a migrazione. Il nostro studio ha un carattere esplorativo e, più che trarre conclusioni, vorrebbe essere uno stimolo ad indagare nell'ambito dell'accomodamento tra varietà nel contesto italiano. Intendiamo dunque proseguire in questa direzione, confrontando il processo di livellamento con quello di adozione di tratti per osservare se sia regolato da fattori simili, in studi di tipo longitudinale con un numero molto più ampio di partecipati, allargando l'indagine (anche acustica) ad altri tratti e varietà di Italiano con diverse caratteristiche linguistiche e sociolinguistiche.

Riferimenti bibliografici

Auer, P., Barden, B., Grosskopf, B. (1998). Subjective and objective parameters determining 'Salience' in long-term dialect accommodation. In *Journal of Sociolinguistics* 2, 163-187.

Auer, P., Hinskens, F., Kerswill, P. (a cura di) (2005). *Dialect change: convergence and divergence in European languages.* Cambridge; New York: Cambridge University Press.

Avano, G., Avesani, C., Vayra, M. (*in stampa*). Gorgia Toscana in terra emiliana: l'accomodamento in fiorentini emigrati in Emilia-Romagna, uno studio sociofonetico. In *Studi e Saggi Linguistici*.

Bigham, D.S. (2010). Mechanisms of accommodation among emerging adults in a university setting. In J*ournal of English Linguistics* 38, 193-210.

Biliotti, F., Calamai, S. (2010). Linguistic opinions and attitudes in Tuscany: verbal guise experiments on the varieties of Arezzo and Florence. In Calamai S., Celata C. & Ciucci L. (a cura di), *Proceedings of the Workshop Sociophonetics, at the crossroads of speech variation, processing and communication*, Pisa: Scuola Normale Superiore, 1-4.

Boersma, P., Weenink, D. (2023). PRAAT: doing phonetics by computer [software]. Versione 6.3.03. https://www.praat.org/

Bross, F. (2019). Using mixed effect models to analyze acceptability rating data. Version 1.0. mimeo.

Calamai, S. (2017). Tuscan between standard and vernacular: a sociophonetic perspective. In Cerruti, M., Crocco, C. & Marzo, S. (a cura di), *Towards a New Standard*. Berlin, Boston: De Gruyter, 213-241.

Calamai, S. (2011). Per una storia della pronuncia degli italiani: opinioni e atteggiamenti intorno alla pronuncia fiorentina. In Nesi, A., Scotti, & M.S., Maraschio, N. (a cura di), *Storia Della Lingua Italiana e Storia Dell'Italia Unita. L'italiano e Lo Stato Nazionale. IX Convegno ASLI* (Firenze 2–4 December 2010), Firenze: Cesati, 175-184.

Castellani, A. (1961). Precisazioni sulla gorgia toscana. In *Centro de estudos filologicos*. 242-262.

Cravens, T.D. (2000). Sociolinguistic subversion of a phonological hierarchy. In *WORD* 51, 1-19.

Cravens, T.D., Giannelli, L. (1995). Relative salience of gender and class in a situation of multiple competing norms. In *Language Variation and Change* 7, 261-285.

De Mauro, T. (2016). Il Nuovo vocabolario di base della lingua italiana. Internazionale. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

Evans, B. (2004). The role of social network in the acquisition of local dialect norms by Appalachian migrants in Ypsilanti, Michigan. In *Language Variation and Change* 16. 153-167.

Evans, B.G., Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. In *j. Acoust. Soc. Am.* 121, 3814-3826.

Foreman, A. (2003). Pretending to be someone you're not: a study of second dialect acquisition in Australia. PhD dissertation, University of Monash.

Galli De' Paratesi, N. (1984). Lingua toscana in bocca ambrosiana: tendenze verso l'italiano standard: un'inchiesta sociolinguistica. Bologna: Il Mulino.

Gallucci, M. (2019). GAMLj: General analyses for linear models. *[jamovi module]*. Retrieved from https://gamlj.github.io/

Giannelli, L., Savoia, L.M. (1978). L'indebolimento consonantico in Toscana I. In *Rivista Italiana di Dialettologia* 2, 25–58.

Giannelli, L. (2000). Toscana, Nuova ed. aggiornata. Ospedaletto (Pisa): Pacini.

Giles, H., Taylor, D.M., Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: some Canadian data. *Lang. Soc.* 2, 177–192.

Kerswill, P. (2013). Koineization. In Chambers, J.K., Schilling, N. (a cura di), *The Handbook of Language Variation and Change*. New York: Blackwell, 519–536.

Labov, W. (2010). *Principles of linguistic change. 1: Internal factors.* Oxford: Blackwell.

Marotta, G. (2001). Non solo spiranti. La 'gorgia toscana' nel parlato di Pisa. In *L'Italia Dialettale* 62, 27–60.

Marotta, G. (2008). Lenition in Tuscan Italian (Gorgia Toscana). In Brandão De Carvalho, J., Scheer, T. & Ségéral, P. (a cura di), *Lenition and Fortition, Studies in Generative Grammar.* Berlino, New York: de Gruyter Mouton, 235–272.

Marotta, G. (2014). New parameters for the sociophonetic indexes: Evidence from the Tuscan varieties of Italian. In Celata, C., Calamai, S. (a cura di), *Studies in Language Variation.* Amsterdam: John Benjamins Publishing Company, 137-168.

Merlo, C. (1950). Gorgia toscana e sostrato etrusco. In *Italica* 27, 253-255.

Milroy, L. (1987). *Language and social networks.* 2nd edition. Oxford: Blackwell.

Nese, N., Meluzzi, C. (2017). Accomodamento ed emergenza di varianti fonetiche: le affricate dentali intermedie a Pavia e Bolzano. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I (a cura di), *Fattori sociali e biologici nella variazione fonetica. Studi AISV 3*, 67-82.

Nespor, M., Vogel, I. (2007). *Prosodic phonology: with a new foreword.* Boston, Berlin: De Gruyter.

Nycz, J. (2015). Second Dialect Acquisition: A Sociophonetic Perspective. L*anguage and Linguistics Compass* 9, 469-482.

Nycz, J. (2016). Awareness and Acquisition of New Dialect Features. In BABEL, A.M. (a cura di), *Awareness and Control in Sociolinguistic Research*. Cambridge: Cambridge University Press, pp. 62-79.

Nycz, J. (2018). Stylistic variation among mobile speakers: Using old and new regional variables to construct complex place identity. In *Lang Var Change* 30, 175-202.

Pardo, J.S., Gibbons, R., Suppes, A. & Krauss, R.M. (2012). Phonetic convergence in college roommates. In *Journal of Phonetics* 40, 190-197.

Piccardi, D., Ardolino, F. (2021). Gaming Variables in linguistic research. Italian Scale Validation and a Minecraft pilot study. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., Schmid, S. (a cura di), *L'individualità del parlante nelle scienze fonetiche: applicazioni tecnologiche e forensi, Studi Aisv 8,* 299–324.

Siegel, J. (2010). *Second Dialect Acquisition.* Cambridge: Cambridge University Press.

Sorianello, P. (2003). Spectral characteristics of voiceless fricative consonants in Florentine Italian, In Solé, M.J., Recasens, D., Romero, J. (a cura di), *Proceedings of the 15th International Congress of Phonetic Sciences.* 3081-3084.

Sorianello, P., Bertinetto, P.M. & Agonigi, m. (2005). Alle sorgenti della variabilità della "Gorgia" fiorentina: un approccio analogico. In Cosi, P. (a cura di), *Convegno Nazionale'AISV - Associazione Italiana di Scienze della Voce, Misura dei Parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici.* Brescia: EDK Editore, 327-362.

The Jamovi Project (2021). jamovi. (Version 1.6) [Computer Software]. Retrieved from https://www.jamovi.org

Trudgill, P. (1986). *Dialects in contact.* Oxford: Blackwell.

Villafaña Dalcher, C. (2008). Consonant weakening in Florentine Italian: A cross-disciplinary approach to gradient and variable sound change. In *Lang Var Change* 20, 275-316.

Vogel, I., 1999. Prosodic phonology. In: Maiden, M. & Parry, M. (a cura di), *The Dialects of Italy*. London and New York: Routledge, 58–68.

DOI: 10.17469/O2111AISV000019

KIM A. GROOTHUIS, GIOVANNI LEO, GIUSEPPE MAGISTRO,
CLAUDIA CROCCO

Prosodic features of pragmatic *quello* in Informal Neapolitan Italian: data from real-world speech[1]

Due to contact with the local dialects, the informal varieties of Italian spoken in Campania feature a pragmatic use of the demonstrative pronoun *quello*, which can be used to double an argumental DP or as an expletive subject. The present paper investigates the prosodic properties of *quello* in a corpus collected by means of *Pyrlato,* a pipeline to extract data from YouTube. 1775 instances of *quello* have been collected, of which 248 were relevant for the analysis. The data show that pragmatic *quello* bears a pitch accent, unlike demonstrative determiner *quello*. Additionally, a prosodic boundary between pragmatic *quello* and the adjacent co-referent DP can be visually and/or auditorily detected, confirming the absence of an intonational contour arching over the two elements. Assuming that the results of our prosodic analysis of *quello* can be extended to its dialect equivalent *chillo*, we formulate a new hypothesis about the origin of pragmatic *chillo*/*quello*.

Keywords: Neapolitan Italian, Neapolitan, *quello*, pragmatic expletives, real-world speech.

1. *Introduction*

Campanian dialects, spoken in the southwest of Italy, show a pragmatic use of the third-person pronoun *chillo*, originally a distal demonstrative (*chillo*, *chella*, *chello* 'that.M/F/N.SG') (Gaeta, 2014; Ledgeway, 2000: 77–78, 2009: 290–294, 915–922, 2010, 2013: 281–282; Maturi, 2002: 225–228, 2023: 80–81; Radtke, 1997: 83–85; Sornicola, 1996; Vitolo, 2006). Despite the fact that Campanian dialects are null-subject languages, this pronoun seems to function as a pragmatic expletive subject of impersonal predicates, in the sense that it does not seem to have an argumental role within the sentence, while adding an interpretative nuance (1):

(1) *Chello* *chiove*.
that.N rain.3SG
'That's because/the fact is that it rains.' (Sornicola, 1996: 336)

Moreover, the same *chillo* is involved in certain topic constructions[2], where it can double an expressed subject (2):

[1] This paper is the result of a close collaboration of all four authors; for official academic purposes only, Kim Groothuis takes responsibility for § 2-3, Giuseppe Magistro for § 4, Giovanni Leo for § 5, and Claudia Crocco for § 6. We assume shared responsibility for § 1 and § 7.

[2] *Chillo* can also double focused subjects, cf. Groothuis (2022) for discussion.

(2) *Chillo* *pateme* *nun* *vò* *capì.*
that.M.SG father=mine NEG want.3SG understand.INF
'My father does not want to understand.' (Sornicola, 1996: 326)

In both (1) and (2), *chillo/chella/chello* is an optional element in the clause in the sense that its absence does not render the clause ungrammatical, unlike canonical expletives such as 'there' or 'it' in non-null-subject languages like English. It does not affect the truth value of a clause either, which means that it serves a discourse function, although it is not easy to pin down which one exactly.

Due to contact, a similar pragmatic use of the demonstrative *quello* (*quello/a*, 'that.M/F.SG', *quelli/e* 'those.M/F.PL'; expletive: (3); topic: (4)) percolates into markedly informal varieties of Campanian Italian (De Blasi and Fanciullo, 2002: 636; Maturi, 2002: 225–228, 2023: 139):

(3) *Quello* *piove.*
that.M rain.3SG
'That's because/the fact is that it rains.' (De Blasi and Fanciullo, 2002: 636)

(4) *E certo, quello* *il brodo di* *gallina* *è* *sostanzioso.*
and sure that the broth of chicken be.3SG substantial
'And certainly, chicken broth is substantial.' (De Filippo, Natale in Casa Cupiello)

While pragmatic *chillo/quello* has received some attention within the syntactic literature, its prosodic properties are practically uncharted, in spite of their most likely relevance for syntactic and pragmatic analysis. The present study sets out to contribute to filling this gap by exploring the prosodic properties of pragmatic *quello* in one of these varieties, i.e. Informal Neapolitan Italian (henceforth INI). The paper is structured as follows: § 2 reviews past literature about the intonation of *quello*; § 3 discusses the main aims and research questions; in § 4 we explain the methodology used for collecting data from real-world speech as found on YouTube; § 5 presents the results; § 6 provides a preliminary analysis of the data, and § 7 concludes the paper.

2. *The intonation of quello: state of the art*

The syntactic literature on pragmatic *quello* makes reference to the intonational pattern of the sentences in which it occurs, but no detailed studies on the matter are available. For instance, it has been claimed that *chillo* in utterances such as (2), when it introduces a new or shifted topic, has "the intonational contour [...] of canonical topic-comment structure and not, for example, that of a rhematic sentence" (Ledgeway, 2010: 262). However, the prosodic properties of such a canonical topic-comment form are not explored in the study. Other contexts with *chillo*, specifically copular clauses as (5), are claimed to be often marked by an exclamative contour (Ledgeway, 2009: 291, 294; Sornicola, 1996: 337; Vitolo, 2006: 65). Also in this

case, however, the prosodic properties of *chillo* and the contour in which it appears are not further explored.

(5) *Chillo* *è* *scemo!*
that.M.SG be.3SG stupid
'He's stupid!'

Moreover, *chillo* and the following determiner phrase (henceforth: DP)[3] would not belong to the same prosodic phrase, which is taken as evidence that the two do not form one syntactic constituent (Ledgeway, 2009: 815, 2010: 260; Sornicola, 1996: 326–327). Similarly, Maturi (2002: 226) states that "*anche quando* quello *sia immediatamente adiacente al SN cui è accordato,* [...]*, una breve pausa o una caduta del contorno intonazionale lo separano dal SN, circostanza che porta ad escludere che in questi casi possa essere analizzato come aggettivo*" (also when *quello* is immediately adjacent to the NP with which it agrees, [...], a short pause or a fall of the intonational contour separates it from the NP, something that leads us to exclude that in these cases it can be analysed as an adjective). Such intriguing observations on a prosodic break (*pausa* or *caduta*) between *chillo/quello* and the DP, or the intonational contour of the sentences in which it occurs, however, are regrettably only based on impressionistic evaluations.

Finally, in the literature the idea has been put forward that prosody plays an important role in the distinction between the different pragmatic uses of *chillo/quello*, cf. e.g., Sornicola (1996: 338) who suggests that "*i fattori prosodici abbiano un qualche influsso sull'intera casistica descritta*" ("prosodic factors have some type of influence on the typology as described"). This same idea is repeated by Vitolo (2006: 69). A systematic study of the interaction between prosody and the syntactic-pragmatic uses of *quello* is, however, still lacking.

In this paper, we aim to take a first step at investigating the prosody of pragmatic pronouns in Southern Italy. The data of choice for a first exploration of a phenomenon typically occurring in dialects or colloquial regional Italian are represented by informal, real-world dialogical productions. At present, we do not dispose of corpora of Neapolitan dialect and Neapolitan Italian featuring informal speech with these characteristics. In fact, speech elicited through task-oriented techniques – such as the CLIPS corpus (www.clips.unina.it), which also contains dialogues from the area of Naples – is not suitable for the exploration we wish to conduct, because pragmatic *quello* occurs only seldom in such a speech style[4]. This suggests that the speaking style of such dialogues is not sufficiently informal to encourage the use of a trait which seems, on the other hand, widespread (Ledgeway,

[3] The determiner phrase is the constituent formed by a D(eterminer) head (e.g. an article or a demonstrative) and an N(oun)Phrase.

[4] Pragmatic *quello* occurs only sporadically in the corpus CLIPS/Naples, despite the strong interactivity of the dialogues included in this corpus. Similarly, a preliminary search in a corpus of interviews with university students from Naples (courtesy of Margherita di Salvo) yielded a total of 23 occurrences of pragmatic *quello* in 12 hours of speech.

2010: 259), and presumably not socially marked in Neapolitan Italian (Groothuis *et al.* in prep.). For this reason, we opted for a different approach to the data collection.

The data examined in this work are extracted from YouTube by means of a pipeline relying on the automatic-generated subtitles (see § 4 for a detailed description of the pipeline). Although social media speech is not necessarily spontaneous in the sense of non-read or non-scripted (Beckman, 1997; Tucker and Mukai, 2023; Warner, 2012), it is normally meant to be addressed to a real audience. Such real-world speech, therefore, can be part of the real-life linguistic experience of the speakers from the area of Naples.

At present, data scraping is only possible for Italian since automatic subtitles in non-standard languages such as Neapolitan are currently not supported by YouTube. Accordingly, the present study focuses on *quello* in the INI as spoken in Naples and the surrounding area. While this choice clearly represents a limitation of the study, it is worth stressing that the analysis of *quello* in INI will also be relevant for future studies on the prosody of *chillo* in Neapolitan. Although a dialect and the regional variety of Italian of the same area are two different languages, it is known from past studies that dialects highly influence the Standard variety on all linguistic levels, also as far as prosody is concerned (Barone, in prep; Cataldo, 2023; Crocco, Gili Fivela, and D'Imperio, 2022; O'Rourke, 2005; Telmon, 1994). So, we assume as a working hypothesis that *quello* and *chillo* also resemble each other intonationally. We leave the study of the intonational of *chillo* in Neapolitan for further research (cf. Groothuis & Crocco 2023, in prep.).

3. *Aims*

The main goal of the study is, thus, to investigate the interaction between the prosody, the syntax, and the pragmatic roles of pronouns such as *quello* in Campania by exploring the prosodic properties of *quello* in a corpus of INI scraped from the Internet. In doing so, we also aim at elaborating the above-mentioned observations originally made on Neapolitan (cf. § 2) concerning:

a. the presence/absence of an overarching contour including pragmatic *quello* and the adjacent co-referent DP (cf. ex. (4)) (RQ1);
b. the presence/absence of detectable prosodic differences when pragmatic *quello* is co-referent with a focalized or topicalized constituent (RQ2);
c. moreover, we intend to compare the intonational contour of the pragmatic uses of *quello* and its use as a non-deictic subject pronoun (RQ3).

The study thereby contributes to the description of the prosody of the Italian linguistic space and especially to the study of informal regional varieties. Since pragmatic *quello* is involved in the expression of topic and focus, the study of its prosodic properties will also contribute to our understanding of the syntax-prosody interface in Italo-Romance.

4. *Methodology*

To collect a corpus of INI, we relied on the *Pyrlato* pipeline (developed for parallel work, cf. Magistro, Crocco, 2023), written in Python, to extract natural data from YouTube videos, mainly movies, TV shows, and comedy sketches. The corpus therefore includes scripted and spontaneous productions in INI.

First, *Pyrlato* uses the module *PyTube* to search for specific keywords (in our case, names of local shows and public figures)[5], resulting in a list of relevant videos. *Pyrlato* then checks that the videos are in the desired language, so as to look for specific strings in the language under scrutiny. The pipeline works for Arabic, Dutch, English, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Romanian, Russian, Spanish, Thai, Turkish, Ukrainian, and Vietnamese. Naturally, we selected Italian for our case study. Next, *Pyrlato* generates an object scraping YouTube's automatic subtitles. These objects contain a Python dictionary with timestamps and the captions of the videos. The captions can be searched to match an exact string or a regular expression. This allows the software to look for more complex patterns such as specific word structures. In our case, a list of the morphological realizations of *quello* (*quello/a/i/e*) was iteratively searched in the captions. If the desired expression is matched, then the respective timestamp is extracted to cut the occurrence from the audio of the video. In our application, the sentence containing the word is extracted in .mp4 and trimmed observing a time-window of 7 seconds. Unfortunately, the timestamps provided by YouTube are not aligned word-by-word but are generally placed between sentential pauses: a cutting window of 7 seconds is then big enough to ensure that *quell** is present in the cut fragment. Finally, .mp4 videos are converted to .mp3 files usable in *Praat* (Boersma, 2001; Boersma and Weenink, 2021).

Data such as those extracted from YouTube are in lossy formats and/or recorded in noisy conditions. For this reason, they are usually discarded in prosodic studies as they do not meet the sound quality standards expected for most acoustic analyses. Such real-world data, however, represent an invaluable source of information concerning the speakers' actual linguistic uses, exactly for their natural character and in spite of their lesser acoustic quality. Of course, this choice has consequences for the type of analyses that can be carried out on the data. In this paper, we focus on the analysis of pitch, which is a relatively robust prosodic feature (Fuchs and Maxwell, 2016). Through auditory and visual inspection in *Praat* (Boersma, 2001; Boersma and Weenink, 2021), we aim at detecting the presence and (if possible) the type of pitch accent occurring on *quello* and on the co-referent DP.

[5] The following keywords have been used to search for videos in the Italian of Naples and the surrounding area: 'Rita De Crescenzo', 'De Filippo', 'Nino D'Angelo', 'Vincenzo Salemme', 'Made In Sud', 'The Jackal', 'Totò film', 'Canale 21', 'Napoli calcio', 'Cucina napoletana', 'Peppe Iodice', 'Cannavacciuolo', 'Massimo Troisi', 'Simone Schettino', 'Teleluna', 'Telegaribaldi', 'Oscar Di Maio', 'Alessandro Siani', 'Gigi D'Alessio', 'Telecapri', 'Arteteca', 'Biagio Izzo', 'Io speriamo me la cavo', and 'L'amica geniale'.

5. *Results*

5.1 Observed types of *quello* in the corpus

In total, 1397 videos (ca. 550h) were scraped, in which 1775 instances of *quello* were found. First, all instances of *quello* have been checked for relevance: utterances produced by speakers from regions other than Campania, and purely deictic uses as a demonstrative adjective or pronoun have not been included in the analysis. A sample of 20 randomly selected instances of *quello* as a demonstrative adjective has been considered in a second phase of the research to compare their characteristics with those of pragmatic *quello*. Finally, sentences in dialect (N=14) have been discarded.

All relevant instances of pragmatic *quello* (N=248) have subsequently been categorized into seven types based on the syntactic context in which they occur. Type I (6)consists of a *quello* reduplicating a preverbal subject (6), whereas type II reduplicates a postverbal subject (7). Types III and IV are pronominal uses of *quello* as the subject of a non-copular (8) and copular verb (9), respectively. Type V refers to the use of *quello* in a cleft sentence (10), where it occurs in initial position, preceding the copular and the clefted constituent, with which it is co-referent(10). In type VI, *quello* is co-referent with a focused DP preceding a copular verb (11)(11). Finally, type VII is the use of *quello* as an expletive without any co-referent DP (12)(12).

(6) *Quella* *mammà* *sta* *in paradiso.* (*La lettera di mammà*)
that.F.SG mom be.3SG in paradise
'Mom is in heaven.'

(7) *Quelli* *rimangono* *impressionati* *i bambini.* (*Bello di papà*)
that.M.PL remain.3PL impressed the children
'The children get creeped out.'

(8) *Quella* *a te* *ti ascolta.* (*Giuro che ti amo*)
that.F.SG to you you=listen.3SG
'She listens to YOU.'

(9) *E* *quell'* *è* *l'umidità.* (*Totò, Peppino e le fanatiche*)
and that.F.SG be.3SG the=humidity
'And that's (because of) the humidity.'

(10) *Quello* *è* *lui ...* *fra* *Martino...* *che* (*I 4 monaci*)
that.M.SG be.3SG him friar Martino that
'That's him... friar Martino... who...'

(11) *Quelle* *due* *so'* (*Arteteca*)
that.F.PL two be.3PL
'They're TWO.'

(12) *Quello* *eravamo* *prima* *un* *gruppo più* *ampio.* (TGR)
that.M.SG be.IPFV.1PL first a group more large
'At first we were a larger group.'

The number of found instances per type is shown in Table 1, column "N".

Next, the instances of *quello* have been inspected visually and auditorily using *Praat* (Boersma, 2001; Boersma and Weenink, 2021). During this stage, we examined the pitch contour on the entire utterances, i.e., by checking both *quello*

and – if present – the co-referent DP. Some instances could not be processed because the audio quality was too low to appreciate reliable pitch contours (N = 53). This was anticipated given the nature of the data (cf. discussion in § 4). To assess the presence of a pitch accent, we relied both on the pitch track observation and on the auditory appreciation of a prominence on *quello*. Note that the unvoiced onset in the stressed syllable of *quello* ['kwel] can make the visual appreciation of the pitch accent more difficult. Beside creating a gap in the pitch track, voiceless obstruents also raise f_0 in the first part of a following vowel in several languages including Italian (Hanson, 2009; Kirby and Ladd, 2016). However, coherently with the regional features characterizing Campania varieties, in our data voiceless stops are often realized as voiced.

Table 1 - *Pitch-accented instances of quello per syntactic type*

SYNTACTIC TYPE	*N*	*PA*	*NON-PA*	*?*	*LOW-Q*
I: *Quello*$_i$ DP$_i$ Pred	30	16	1	5	8
II: *Quello*$_i$ Pred DP$_i$	6	5	0	1	0
III: *Quello*$_{subj}$	99	62	2	15	20
IV: *Quello*$_i$ Cop DP$_i$	102	61	5	11	25
V: *Quello*$_i$ cleft DP$_i$	6	4	0	2	0
VI: *Quello*$_i$ DP$_i$ Cop	1	1	0	0	0
VII: *Quello*$_{Expl}$	4	3	0	1	0
Total	248	151	8	35	53

5.1.1 *Quello* as weak deictic subject pronoun

Before we proceed to the presentation of the results, a clarification is needed concerning the inclusion of *quello* as a weak deictic subject pronoun (Type III and IV, examples 8-9) in the corpus[6].

The use of *quello* as a weakly-deictic subject pronoun may seem *prima facie* unrelated to the other cases examined in this study for two reasons: first, the cases included in the analysis seem identical to other possible instances of *quello* being used as a deictic pronominal subject in standard Italian (e.g. "*trovi brutto anche il cappello?*" "*no, quello mi piace*"; "*quella è la casa dei miei sogni*"); second, they look different from the other pragmatic cases where *quello* is co-referent with a DP belonging to the same sentence (e.g. "*quello tuo figlio sta sempre da solo*"). However, it can be argued that the cases of *quello* as a weakly-deictic subject pronoun included in the analysis exhibit a specific pragmatic nuance that justifies their inclusion in the corpus. In this subsection, we will briefly explain why we think this use of *quello* is pragmatically marked.

[6] By defining such cases as weakly-deictic, we wish to stress that, in many cases, although the pronoun does not have a clear deictic function, it can be argued that pronoun has some kind of discourse reference. In fact, the distinction between deictic and non-deictic *quello* is often hard to make.

In cases such as "*quella a te ti ascolta*", *quella* has a weak deictic function, as it simply refers to an identifiable subject. In a less regionally flavoured version of the same sentence, *quella* could be replaced by *lei* ("she"), without changing the meaning of the sentence: *"lei a te ti ascolta"*. The use of *quello/quella* as a subject pronoun is likely to be a regional use rooted in the dialect: the expansion in 18th century of *chillo/chella* in Neapolitan as a subject pronoun at the expense of the subject pronoun *isso/essa* (i.e. without a clear deictic function) has been pointed out by Ledgeway (2009: 276-280, 286-290). Similarly, the Neapolitan speakers of Italian can use both *quello/quella* and *lui/lei* as subject pronouns (both sentences "*quella a te ti ascolta*" and "*lei a te ti ascolta*" are acceptable in the variety). While the alternation between *quello/quella* and *lui/lei* can possibly be affected by the speech style – regional features tend indeed to emerge in informal registers (Berruto, 1987 [2012]) – it can be argued that the expansion of *quello* as a subject pronoun is pragmatically motivated.

Moreover, the two variants of the subject pronoun (*quello/quella* vs *lui/lei*) are not freely interchangeable but have different syntactic properties, as in Neapolitan (Ledgeway, 2009: 290). In particular, two syntactic differences can be noted. First, only *isso/essa* can have a local antecedent (i.e., be co-referent with a DP in the same clause), cf. (13) where only *essa* is grammatical. Second, in an embedded clause, *isso* can refer anaphorically to the main clause subject or have a disjunctive interpretation; *chillo*, on the other hand, can only be interpreted disjunctively (see example (14)):

(13) *Vicenza$_i$ s'è messa a vennere 'o ccafè pur'*
Vicenza REFL=be.3SG put to sell.INF the coffee also
*essa$_i$/*chella$_i$.*
she/that.F.SG
'Vicenza started to sell coffee also herself.' (Ledgeway, 2009: 920)

(14) *Ciro$_i$ ha itto ca isso$_{i+j}$/ chillo$_{*i/j}$ non capesce*
Ciro have.3SG said that he that.M.SG NEG understand.3SG
niente.
nothing
'Ciro has said that he does not understand anything.' (Ledgeway, 2009: 920)

The same can be noticed for regional Italian *quello*, which differs from *lui/lei* along the same lines:

(15) *Vincenza s'è messa a vendere il caffè pure*
Vicenza REFL=be.3SG put to sell.INF the coffee also
*lei /*quella.*
she /that.F.SG
'Vicenza started to sell coffee also herself.'

(16) *Ciro ha detto che lui$_{i+j}$/ quello$_{*i/j}$ non capisce*
Ciro have.3SG said that he that.M.SG NEG understand.3SG
niente.
nothing
'Ciro has said that he does not understand anything.'

Also more generally, the use of a subject pronoun in a null subject language is automatically marked. In unmarked situations, the more natural choice would be to use no pronoun (cf. the Avoid pronoun principle, Chomsky, 1981). A subject pronoun is used when it has a marked pragmatic interpretation (be it topic or focus). This is in line with the Division of labour principle as proposed by Horn (1984), according to which the use of a marked expression when a corresponding unmarked form is available tends to be interpreted as conveying a marked message.

In other words, speakers can use *quello* as a subject pronoun to indicate that the subject has a degree of topicality, as it is present in the discourse, though not necessarily in the immediate context in which the interaction takes place. This is unsurprising given the high degree of topicality of pronouns and the frequent overlap of the categories topic and subject (Chafe, 1976). Therefore, the case of *quello* as a subject pronoun shows that, in certain contexts, the boundaries between an unmarked pronoun and a topicality marker are blurred.

5.2 Prosodic inspection of *quello* per type

From the observation of our corpus, *quello* appears to be marked by a high or rising f_0 movement that can be interpreted as a pitch accent in 152 out of 248 occurrences across the different syntactic-pragmatic contexts. In other cases, the presence of a pitch accent could not be determined for sure (N=35). The results of this inspection can be seen in Table 1, where PA = presence of a pitch accent on *quello*, NON-PA = absence of pitch accent, ? = uncertainty about the presence of the pitch accent and LOW-Q indicates those data excluded in the analysis because of the low quality of the audio. In the remainder of this section, we will discuss the results more in detail per type of *quello*.

When occurring with a co-referent DP (Types I and II), both *quello* and the co-referent DP seem to carry a pitch accent (cf. Fig. 1, 2, 3 and 4). Out of 36 instances, a pitch accent on *quello* could be individuated in 21 cases. There were 14 cases where the presence of a pitch accent was uncertain or could not be determined due to low signal quality. Only in one example, there was no pitch accent on *quello.*

In the cases where *quello* and the co-referent DP are both preverbal (Type I), they mostly belong to the same major prosodic phrase, although a degree of disjuncture between *quello* and the DP can be detected, hinting to a minor prosodic break separating the two elements. Additionally, *quello* and the co-referent DP can be separated by a pause (Fig. 4), a pitch drop (Fig. 3) or by intervening material (Fig. 2). Such cases strengthen the observation that *quello* and the co-referent DP are not necessarily included in the same phrase.

Figure 1 - *Example of quello + preverbal DP (Type I, "That Mom is in paradise")*

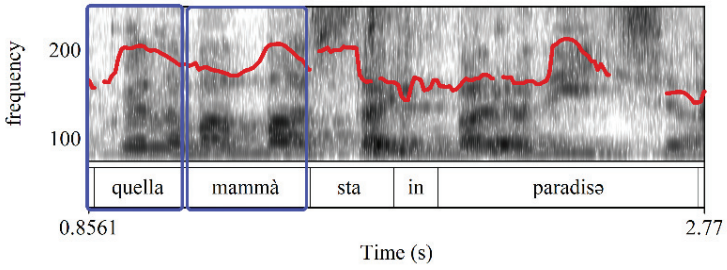


Figure 2 - *Example of quello + preverbal DP with intervening material (Type I, "That sometimes loneliness plays nasty tricks on you")*

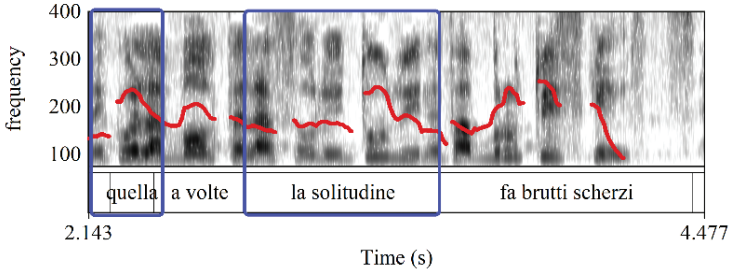


Figure 3 - *Example of quello + preverbal DP with pitch drop (Type I, "That uncle Pasquale has just arrived from Switzerland")*

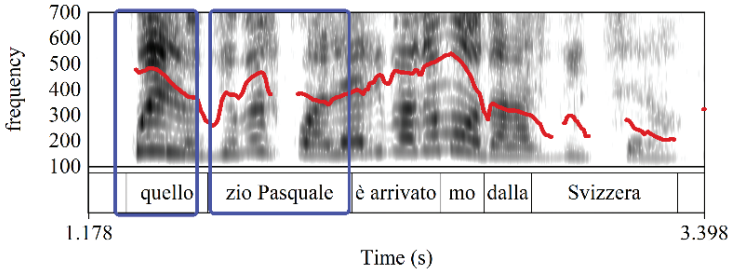
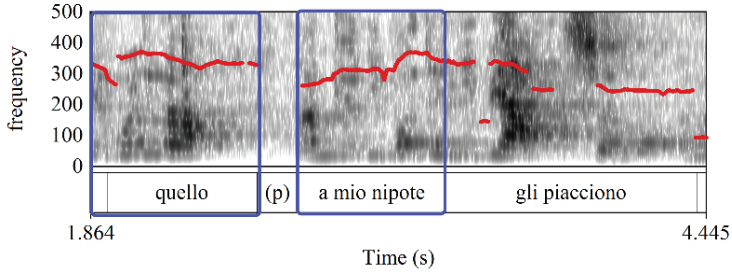


Figure 4 - *Example of quello + preverbal DP with pause, indicated in transcription as "p" (Type I, "That my nephew likes them")*



For type II, where the co-referent DP is postverbal, a pitch accent can mostly be detected on *quello* (see Fig. 5 and 6). The co-referent DP can be in post-focal position (Fig. 5).

Figure 5 - *Example of quello + postverbal DP (Type II, "Those they get creeped out, the children")*

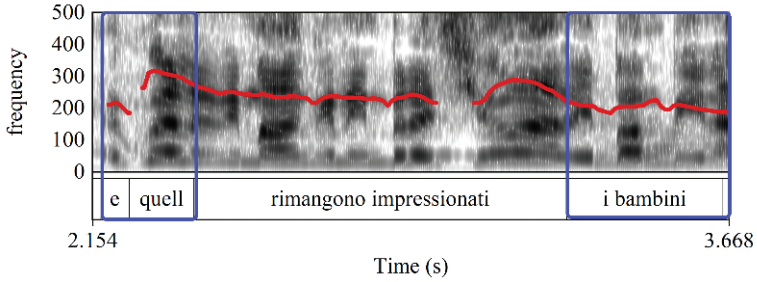
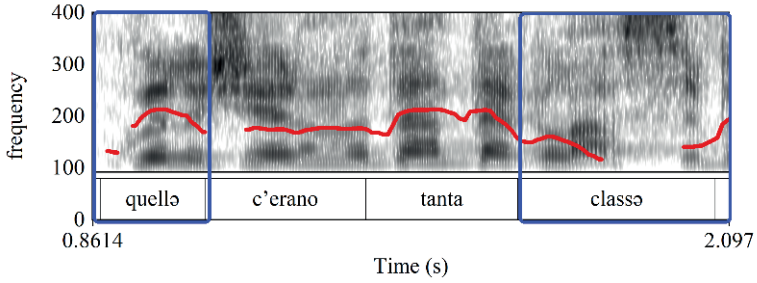


Figure 6 - *Example of quello + postverbal DP (Type II, "Those there were many classes.")*



Overall, in those cases where *quello* is associated with a co-referential DP, both items seem to be pitch accented, although the properties of the accented DP depend on its focal or post-focal position, as expected in this variety (D'Imperio 2001). The case presented in Fig. 5 illustrates a post-focal DP *i bambini* ("the children"), functioning as sentence topic, which is realized with a reduced or downstepped pitch movement (D'Imperio, 2001). In Fig. 6 the DP *tanta classe* ("so many classes/classrooms") is focused and carries the nuclear accent.

When used as a subject pronoun, both of copular and non-copular verbs (types III and IV), *quello* carries a pitch accent of its own, as shown in Fig. 7 and 8. The shape of f_0 movement is similar to the one we observed in the previous types, i.e., a rise / high.

Figure 7 - *Example of quello as a subject pronoun DP, non-copular verb (Type III, "That he will arrest us")*

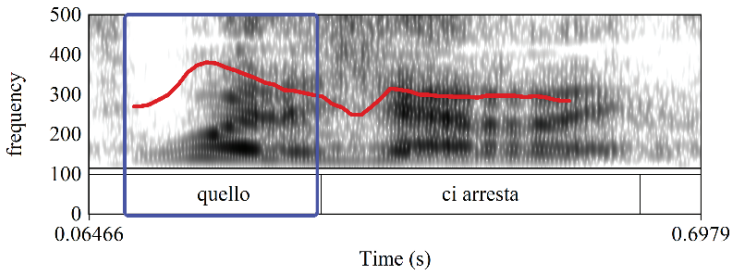
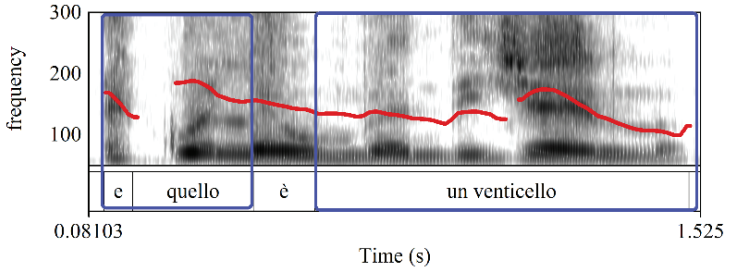


Figure 8 - *Example of quello as a subject pronoun DP, copular verb (Type IV, "That it is a little breeze")*



Also in types V and VI, *quello* exhibits a pitch accent. We can see this in Fig. 9 for type V, where a focused DP occurs before the copula, and in Fig. 10 for type VI, where the focused DP is a clefted constituent.

Like the other types, expletive *quello* is also marked by a high/rising pitch accent. An example can be seen in Fig. 11. In this case, *quello* is not co-referent with an argument within the sentences but refers to a proposition present in the discourse (cf. Ledgeway, 2010: 286-288; Groothuis, Crocco, in prep.).

Figure 9 - *Example of quello with a focused DP (Type V, "Those they are TWO")*

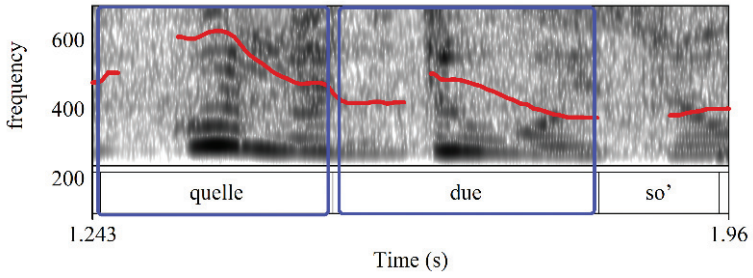


Figure 10 - *Example of quello with a clefted DP (Type VI, "That it is the director let's say that tends to...")*

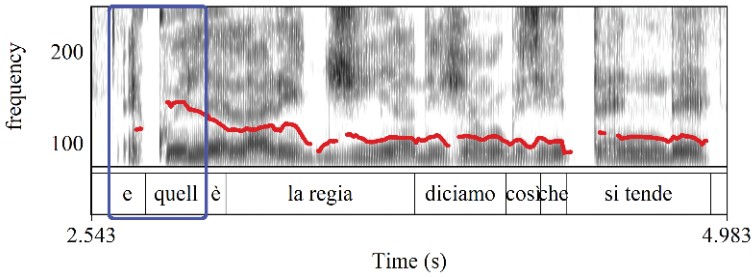
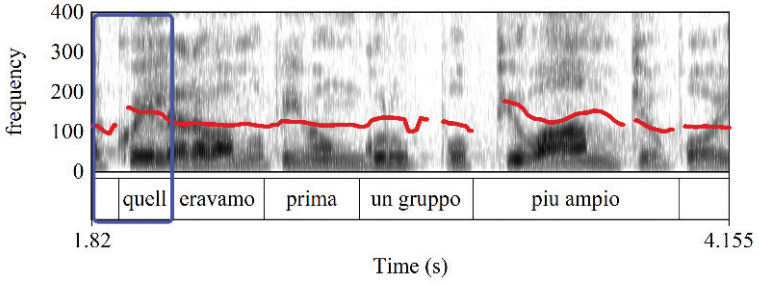


Figure 11 - *Example of expletive quello (Type VII, "That before we were a bigger group")*



As far as can be observed from the corpus, no detectable prosodic differences emerge in the realization of *quello* according to the pragmatic context in which the pronoun appears. *Quello* can indeed refer to a topical or focal DP, but the discourse category of the co-referent DP does not seem to affect its realization. However, given the low acoustic quality of the data at our disposal, more targeted investigations are necessary to confirm this observation. At the same time, the accent marking the DP does vary according to its (pre-, post-, or focal) position in the utterance, as expected based on previous evidence on Neapolitan Italian (D'Imperio, 2002).

6. *Discussion: demonstrative quello vs pragmatic quello*

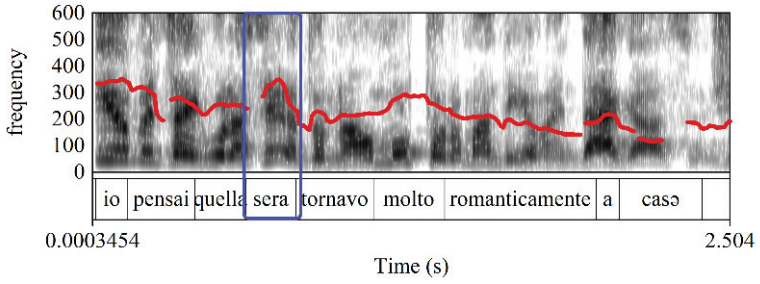
Through *Pyrlato*, we were able to extract ca. 200 occurrences of pragmatic *quello* that could undergo a prosodic analysis out of a total amount of 1729 items. While this may seem a very small amount of usable data, the picture changes if we consider that spontaneous utterances containing a pragmatic *quello* are difficult to obtain[7]. With this in mind, automatic scraping of internet data represents a valuable technique to gather informal data suitable for the investigation of phenomena that are otherwise difficult to observe.

[7] Cf. fn. 4.

With regard to the first research question (RQ1) about the presence/absence of an overarching contour including pragmatic *quello* and the adjacent co-referent DP (Type I), the data indicate that both elements are pitch-accented and can be separated by an audible boundary. However, the strength and the actual realization of such a boundary, e.g., as a tonal event, can vary in our data. Occasionally, *quello* and the following DP can be separated by intervening material (Fig. 2) as well as by a pause (Fig. 4).

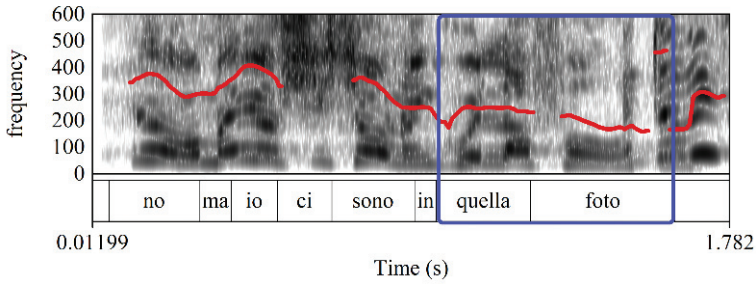
Such observations concerning Type I *quello* can be integrated with the results of a comparison with the context in which *quello* appears as a demonstrative determiner. In many of the occurrences, collected by means of *Pyrlato*, *quello* is a demonstrative determiner (i.e., *quello* with an adjectival value, as in *quella casa* "that house" or *quel ragazzo* "that boy"), as expected in Standard Italian. We inspected a sample of 20 cases of *quello* used as a demonstrative determiner. In these instances, the modified noun rather than the pronoun *quello* seems to be the bearer of the pitch accent (Fig. 12 and 13)[8]. Furthermore, no audible boundary can be detected between *quello* and the modified noun, that seem therefore included in the same prosodic phrase. With respect to Type I *quello*, demonstrative determiners seem to feature a different prosodic organization, characterized by the "overarching contour" hypothesized a.o. by Ledgeway (2010). By and large, our results for Type I *quello* converge with the observations put forward by Sornicola (1996), Maturi (2002), and Ledgeway (2010) on Neapolitan *chillo*. This result also provides support to the related hypothesis that the two items do not form one syntactic constituent, unlike determiner *quello* and the modified noun.

Figure 12 - *Example of demonstrative determiner quello ("I thought that that night I would go back home very romantically")*



[8] Only in one case (*quelle caratteristiche*, 'those features') both *quello* and the following noun are clearly accented.

Figure 13 - *Example of demonstrative determiner quello ("No, but I am there in that picture")*



The data gathered through *Pyrlato* indicate that the prosodic properties of the co-referential DP vary according to its position with respect to the focus, thereby confirming the findings of previous research on the Neapolitan variety of Italian (D'Imperio, 2001, 2002). As far as it is possible to observe in the data examined in this paper, *quello* does not present detectable prosodic differences when it is co-referent with a topicalized or focalized constituent (RQ2). However, we are not able to further elaborate on this point due to the signal quality of the collected material. Therefore, as the prosodic realization of *quello* cannot be adequately explored in the available dataset, we leave this issue open for a different work based on more controlled data (Author, Author in prep.).

We also intended to compare the intonational contour of the pragmatic uses of *quello* with its use as a weakly-deictic subject pronoun (RQ3). As discussed in § 5.1.1, the boundaries between an unmarked pronoun and a topicality marker are blurred in the case of non-deictic pronominal *quello* (Type III-IV)[9]. Also at the prosodic level, we do not see a clear difference between *quello* as a weak deictic subject pronoun and the pragmatic uses of *quello*. In fact, we notice that *quello* in type I-II, which has a clear pragmatic function, and *quello* in types III-IV, in which it could be ambiguously interpreted as a topic and a subject, feature a high/rising pitch accent. No further prosodic property is found to have a distinctive role across these categories. Instead, the relevant prosodic distinction seems to be between the demonstrative determiner *quello* (as in *quella sera*, "that night") and the other uses examined in this study: while the demonstrative determiner does not normally bear a pitch accent, it tends to be pitch accented when it occurs with a pragmatic or pronominal value.

On the assumption that the results of our prosodic analysis of *quello* can be extended to its dialectal equivalent *chillo*, we formulate a new hypothesis about the origin of the pragmatic use of *chillo/quello*, namely that the pitch-accented demonstrative pronoun *chillo/quello* may have provided prosodic cues for the functional reanalysis of pronominal *quello* as a pragmatic marker. When used as a subject pronoun, *quello* must bear a certain degree of discourse salience, because

[9] Note, however, that *quello* can be co-referent with a focused element (type V and VI), suggesting that this pronoun does not mark topicality per se.

otherwise a null pronoun/pro-drop would be preferred (see § 5.1). Because of its salience at the level of the discourse, the pronoun interpretation is ambiguous between a regular preverbal subject pronoun in the IP-domain and a pragmatic category (i.e., a topic or a focus element) encoded in the left periphery of the clause (Ledgeway, 2010; Rizzi, 1997; Rizzi and Bocci, 2017). We propose that it is exactly this ambiguous context which has permitted the reanalysis of *quello/chillo* as a left-peripheral item.

Although difficult to pin down, the pragmatic function of *quello* can be inferred from its association with co-referent topical or focal elements, as in examples (6), (7), (8), (9) and (10), (11) respectively. Moreover, it can refer to a situation in the common ground when occurring in a thetic sentence (cf. (12) and Fig. 11). In such pragmatically marked contexts, *quello* is also pitch accented and followed by a prosodic boundary, suggesting that it may maintain the prosodic features of the discourse-salient pronoun along reanalysis. The proposed origin of the pragmatic use of the pronoun *chillo/quello* rather than its use as demonstrative determiner is also in line with the wider tendency in Romance of third person pronouns developing into pragmatic expletives (Corr, 2017; Hinzelin, 2009). Parallel examples can be found in Portuguese (*ele*, Carrilho, 2008), Dominican Spanish (*ello,* Hinzelin and Kaiser, 2007)) and Balearic Catalan (*ell,* Bartra-Kaufmann, 2011; Remberger and Hinzelin, 2009). On the other hand, *quello/chillo* as a demonstrative determiner does not show the same ambiguity, nor is pitch accented or phrased apart from the co-referent DP.

The development of *chillo* (< Lat. ECCU ILLU(M)), which should form the basis for the pragmatic use of *quello* examined in this article, can thus be schematized as follows:

(17) demonstrative determiner > subject pronoun > pragmatic left-peripheral pronoun.

7. *Conclusions*

The results presented in this article show that social media are a valuable source of spoken data, which can offer new insights into linguistic analysis despite the limitations presented by the low acoustic quality of the signal. By using *Pyrlato*, a novel pipeline for the extraction of spoken data from the internet (cf. Magistro, Crocco, 2023), we collected a corpus of INI featuring real-world speech from talk-shows, movies, theatre, interviews etc. This data collection allowed us to explore the prosodic properties of pragmatic *quello*, a regional feature typical of informal dialogical speech which seems otherwise difficult to elicit through widely used paradigms for prosodic and syntactic research (task-oriented dialogues, Discourse Completion Task, questionnaires, interviews, etc.).

Through auditory and visual inspection of the scraped data, it was shown that *quello* is frequently marked with a high/rising pitch accent when used as a pronoun or as a pragmatic marker. The data support the hypothesis put forward in previous

studies concerning the different prosodic facets characterizing demonstrative determiner *quello* and pragmatic *quello*, in terms of prominence and phrasing. This has led us to formulate a new hypothesis about the origin of the pragmatic uses of *quello/chillo*, namely that it derives from the use of *chillo* as a pronoun. Although real-world data scraped from the internet have a high potential for prosodic and linguistic research, they also have limitations. These have emerged in the case of RQ2. In fact, social media data did not help in casting light on finer differences possibly differentiating between *quello* co-referent with a focal and *quello* co-referent with a topical DP. This limitation highlights the importance of data triangulation in prosodic research (Vanrell, Feldhausen, & Lluïsa Astruc, 2018).

Of course, the assumed parallel between dialectal prosody and the prosody of *quello* in INI deserves a more in-depth investigation. This matter will be addressed in future work on *chillo/quello* specifically (cf. Groothuis, Crocco, 2023, in prep.), as well as more generally for Campania (Leo in prep.). However, we believe that the results of the prosodic inspection of *quello* reported in this article provide us with valuable first insights into the prosodic properties and the historical developments of pragmatic pronouns in varieties spoken in Campania.

Acknowledgements

The authors wish to thank the audience of the AISV conference in Lecce for helpful discussions on the topic, as well as two anonymous reviewers for their valuable comments and remarks. All remaining errors are the authors' own responsibility. This work was financially supported by the Research Foundation – Flanders (FWO), project numbers 12C0723N (Kim A. Groothuis), G010122N (Claudia Crocco & Giovanni Leo), and G004519N (Anne Breitbarth, Claudia Crocco & Giuseppe Magistro).

References

Barone, M. (in prep). *Intonation in Pescara: The description of two languages in contact.* PhD dissertation, Università degli studi di Lecce/Universitat Pompeu Fabra.

Bartra-Kaufmann, A. (2011). Recycled neuter expletive pronouns and the nature of the left periphery: *Ell* and relatives revisited. In *Catalan Journal of Linguistics*, 10, 185–219. https://doi.org/10.5565/rev/catjl.37

Beckman, M.E. (1997). A typology of spontaneous speech. In Sagisaka, Y., Campbell, N. & N. Higuchi (a cura di), *Computing prosody: Computational models for processing spontaneous speech*. New York: Springer, 7-26.

Berruto, G. (1987 [2012]). *Sociolinguistica dell'italiano contemporaneo* (2nd ed.). Rome: Carocci.

Boersma, P. (2001). *Praat*, a system for doing phonetics by computer. In *Glot. Int.*, 5(9), 341–345.

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer [Computer program]. Version 6.1.40. https://www.praat.org/

Carrilho, E. (2008). Beyond doubling: Overt expletives in European Portuguese dialects. In Barbiers, S., Koeneman, O., Lekakou, M. & Ham, van der, M. (a cura di) *Microvariation in syntactic doubling*. Bringley: Emerald group, 301–323. https://doi.org/10.1163/9781848550216_012

Cataldo, V. (2023). *The intonation of closely related varieties. The case of Campania Italian*. PhD thesis, Università degli Studi di Salerno/Ghent University.

Chafe, W.L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View in Subject and Topic. In Li, C.N. (a cura di), *Subject and topics*. New York: Academic Press, 25–55.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Dordrecht: Foris.

Corr, A. (2017). The grammaticalization of epistemicity in Ibero-Romance: Alike processes, unlike outcomes. In *Journal of Historical Linguistics*, 7(1), 48–76. https://doi.org/10.1075/jhl.7.1-2.03cor

Crocco, C., Gili Fivela, B., & D'Imperio, M. (2022). *Comparing prosody of Italian varieties and dialects: Data from Neapolitan*. In Frota, S., Cruz, M., & Vigário, M. (a cura di), Speech Prosody 2022, Lisbon, 23-26 May 2022, 140–144. https://doi.org/10.21437/SpeechProsody.2022-29

De Blasi, N., Fanciullo, F. (2002). La Campania. In Cortelazzo, M., Marcato, C. (a cura di), *I dialetti italiani. Storia, struttura, uso*. Turin: UTET, 628–678.

D'Imperio, M. (2001). Focus and tonal structure in Neapolitan Italian. In *Speech Communication*, 33(4), 339–356. https://doi.org/10.1016/S0167-6393(00)00064-9

D'Imperio, M. (2002). Italian intonation: An overview and some questions. In *Probus*, 14(1). https://doi.org/10.1515/prbs.2002.005

Fuchs, R., Maxwell, O. (2016). The Effects of mp3 Compression on Acoustic Measurements of Fundamental Frequency and Pitch Range. In Barnes, J., Brugos, A., Shattuck-Hufnagel, S. & Veilleux, N. (a cura di.) *Speech Prosody 2016*, 523–527. ISCA. https://doi.org/10.21437/SpeechProsody.2016-107

Gaeta, L. (2014). Soggetto o topic? Chillo 'o fatto è nire, nire. In Cugno, F., Mantovani, L., Rivoira, M. & Specchia, M.S. (a cura di), *Studi linguisitici in onore di Lorenzo Massobrio*. Turin: Istituto dell'Atlante Linguistico Italiano, 465–479.

Groothuis, K.A. (2022). Reconsidering the pragmatic use of chillo in Campania. Paper presented at *Going Romance 2022*, Universitat Autònoma Barcelona.

Groothuis, K.A., Crocco, C. (2023). The prosodic features of pragmatic *chillo/quello* in Neapolitan and Campanian Italian: implications for syntax and diachrony. Paper presented at the 17th Cambridge Italian Dialects Syntax Morphology Meeting, University of Zürich.

Groothuis, K.A., Crocco, C. (in prep.). The prosodic features of pragmatic *chillo/quello* in Neapolitan and Campanian Italian: implications for syntax and diachrony. Ms, Ghent University.

Groothuis, K.A., Crocco, C., Di Salvo, M. & Matrisciano-Mayerhofer, S. (2024). Pragmatic "quello" from dialect to regional standard. Paper presented at the Romance Grammars, context and contact 2024, University of Birmingham.

Hanson, H.M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. In *The Journal of the Acoustical Society of America*, 125(1), 425–441. https://doi.org/10.1121/1.3021306

Hinzelin, M.-O. (2009). Neuter pronouns in Ibero-Romance: Discourse reference, expletives and beyond. In Kaiser, G.A., Remberger, E.-M. (a cura di), *Proceedings of the Workshop «Null-subjects, expletives, and locatives in Romance»*. Konstanz: Universität Konstanz, 1–25.

Hinzelin, M.-O., Kaiser, G.A. (2007). El pronombre ello en el léxico del español dominicano. In Mihatsch, W., Sokol, M. (a cura di), *Language contact and language change in the Caribbean and beyond*. Frankfurt am Main: Lang, 171-188.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-and R-based implicature. In Schriffin, D. (a cura di), *Meaning, form, and use in context: Linguistic applications*. Washington, DC: Georgetown University Press, 11–42.

Kirby, J.P., Ladd, D.R. (2016). Effects of obstruent voicing on vowel *F* 0: Evidence from "true voicing" languages. In *The Journal of the Acoustical Society of America*, 140(4), 2400–2411. https://doi.org/10.1121/1.4962445

Ledgeway, A.N. (2000). *A comparative syntax of the dialects of Southern Italy: A minimalist approach*. Oxford: Wiley-Blackwell.

Ledgeway, A.N. (2009). *Grammatica diacronica del napoletano*. Tübingen: Niemeyer.

Ledgeway, A.N. (2010). Subject licensing in CP: the Neapolitan double-subject Construction. In Benincà, P., Munaro, N. (a cura di), *Mapping the left periphery*. Oxford: Oxford University Press, 257–296. https://doi.org/10.1093/acprof:oso/9780199740376.003.0009

Ledgeway, A.N. (2013). Testing linguistic theory and variation to their limits: The case of Romance. In *Corpus,* 12, 271–327. https://doi.org/10.4000/corpus.2408

leo, G. (in prep.). Intonation and rhythm of dialects and Italian. The case of Campania. PhD thesis, Ghent University.

Magistro, G., Crocco, C., (2023). Pyrlato: a novel methodology to collect real-world acoustic data. In *Estudios de Fonética Experimental*, 35 (ahead of print).

Maturi, P. (2002). *Dialetti e substandardizzazione nel Sannio Beneventano*. Frankfurt: Lang.

Maturi, P. (2023). *Napoli e la Campania*. Bologna: Il mulino.

O'Rourke, E. (2005). *Intonation and language contact: A case study of two varieties of Peruvian Spanish*. PhD dissertation, University of Illinois at Urbana-Champaign.

Radtke, E. (1997). *I dialetti della Campania, edizione italiana a cura dell'autore, con la collaborazione di Paolo Di Giovine e Franco Fanciullo*. Roma: Il Calamo.

Remberger, E.-M., Hinzelin, M.-O. (2009). *Expletive vs. Discourse marker: Sardinian* ello *and Balearic Catalan* ell. Talk given at the 38th Romance Linguistics Seminar, Trinity Hall, Cambridge. Trinity Hall, Cambridge.

Rizzi, L. (1997). The fine structure of the left periphery. In Haegeman, L.M.V. (a cura di), *Elements of grammar*. Dordrecht: Kluwer, 281–337. https://doi.org/10.1007/978-94-011-5420-8_7

Rizzi, L., Bocci, G. (2017). Left periphery of the clause. Primarily illustrated for Italian. In Everaert, M.B.H., Riemsdijk, van, H. (a cura di), *The Wiley Blackwell Companion to Syntax* (2nd ed.). Oxford: John Wiley & Sons, 1–30. https://doi.org/10.1002/9781118358733.wbsyncom104

Sornicola, R. (1996). Alcune strutture con pronome espletivo nei dialetti italiani meridionali. Benincà, P., Cinque, G., De Mauro, T., & Vincent, N. (a cura di), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*. Rome: Bulzoni editore, 323–340.

Telmon, T. (1994). Gli italiani regionali contemporanei. In Serianni, L., Trifone, P. (a cura di), *Storia della lingua italiana*: *Vol. 3: Le altre lingue*. Turin: Einaudi, 597–626.

Tucker, B.V., Mukai, Y. (2023). *Spontaneous Speech* (1st ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108943024

Vanrell, M.D.M., Feldhausen, I., & Lluïsa Astruc. (2018). The Discourse Completion Task in Romance prosody research: Status quo and outlook. In Feldhausen, I., Reich, U., Vanrell, M.D.M. (a cura di), *Methods in prosody: a Romance language perspective*. Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.1441345

Vitolo, G. (2006). L'uso del dimostrativo espletivo in contesto di frasi principali e dipendenti nella parlata salernitana. *Quaderni del Dipartimento di Linguistica - Università di Firenze*, 16, 61–77.

Warner, N. (2012). Methods for studying spontaneous speech. In Cohn, A., Fougeron, C. & Huffman, M. (a cura di), *The Oxford handbook of laboratory phonology*. Oxford: Oxford University Press, 621–633.

DOI: 10.17469/O2111AISV000020

VINCENZO SCOTTI, ROBERTO TEDESCO

On the Role of Dialogue Context in Predicting Speaking Style

Text-to-Speech (TTS) synthesis is a problem almost as old as Natural Language Processing (NLP). The focus of this problem is on creating tools capable of generating a voice uttering a given text. Deep learning-powered solutions try to go beyond mere speech generation from the text: newer models try to factorise the probability predicted by these generative models to condition it on various aspects: speaker's voice, speaking style, or prosody. In this work, we focused on predicting the speaking style from the given text inside a conversation, for the application to conversational agents and chatbots. To this end, we developed and trained a neural network module working as a connector between the textual component of a chatbot (i.e., the neural language model for dialogue understanding and generation) and the speech synthesis component of a chatbot (i.e., the neural TTS synthesis model with speaking style conditioning).

Keywords: Natural Language Processing, Deep learning, Expressive speech synthesis, Global Style Token.

1. *Introduction*

Large deep learning-based generative models have become a standard tool for Natural Language Processing (NLP). They can be used for text understanding and generation (Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, Liu, 2020; Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal Neelakantan Shyam Sastry, Askell, Agarwal, Herbert-voss, Krueger Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, Mccandlish, Radford, Sutskever, Amodei., 2020), as well as for speech recognition and synthesis (Tan, Qin, Soong, Liu, 2021; Radford, Kim, Xu, Brockman, Mcleavey, Sutskever, 2022). Recent advances have shown how the text-based models can be adopted to build incredibly versatile chatbots (i.e., conversational agents, OpenAI, 2022; Google, 2023).

In this work, we are interested in developing a solution to be adopted by empathetic chatbots. These chatbots simulate empathy to the end of being perceived as human like by the users. They find applicability in applications like automatic therapy, counselling, and coaching. Research on empathetic chatbots showed that embodying a conversational agent though voice or an avatar improves the pro-social attitude of the user towards the agent, helping the human-like perception (Paiva, Leite, Boukricha, Wachsmuth, 2017).

For this scope, we developed our Dialogue Global Style Token (DGST) module. This module builds on top of state-of-the-art approaches for Expressive Speech Synthesis that use a latent code to condition the generated speech towards a specific

style. This latent code is called Global Style Token (GST, Wang, Stanton, Zhang, Skerry-Ryan, Battenberg, Shor, Xiao, Jia, Ren, Saurous, 2018).

We divide the rest of this paper into the following sections. In § 2 we present the state of the art on neural network-based TTS synthesis. In § 3 we describe the methodology we used to infer speaking style from text using dialogue language models in the DGST model. In § 4 we describe the experimental approach we followed to search for the best DGST configuration to predict speaking style from text, the dialogue data we used in our experiments and the results obtained from the different configurations. Finally, in § 5 we summarise our contribution and suggest possible future extensions.

2. *State of the art on neural TTS*

In this section we present the current state of the art for TTS synthesis, with emphasis on the techniques for expressive speech synthesis. Neural network-based TTS use a pipeline divided into two main components: acoustic model (also called spectrogram predictor) and vocoder (Jurafsky, Martin, 2022; Tan *et al.*, 2021). The former component takes care of converting a sequence of graphemes[1] or phonemes[2] into a (Mel-) spectrogram; the latter component takes care of converting the Mel-spectrogram into a raw waveform, concluding the synthesis process.

2.1 Neural network-based speech synthesis

The acoustic models for Mel-spectrogram prediction are commonly built using a Seq2Seq transducer architecture (referring to Fig. 1, the bordeaux and green parts are respectively the input phoneme/grapheme sequence and the output (Mel-) spectrogram). These models are mainly composed by an encoder of the input text and a decoder that generates the output speech. Often the architecture is extended to allow for some form of conditioning on the generated speech. These networks are trained to minimise the mean squared error (MSE) between the reference and predicted spectrogram.

The encoder takes as input a sequence of graphemes or phonemes. Some models even allow using both encodings (Valle, Li, Prenger, Catanzaro, 2020). The symbols in the input sequence are considered orthogonal symbols and are used to fetch embedding vectors that are processed by the encoder. The encoder usually leverages a bi-directional approach to find a latent representation of the input that can be used to condition the decoder to generate the correct spectrogram. The sequence of latent vectors (one for each phoneme or grapheme in the input) in output from the

[1] In this context we define a *grapheme* as the "the smallest meaningful contrastive unit in a writing system". In other words, it is a written symbol that allows distinguishing a *minimal pair* (e.g., the graphemes h and m allow to distinguish the two words in the minimal pair [house, mouse]).

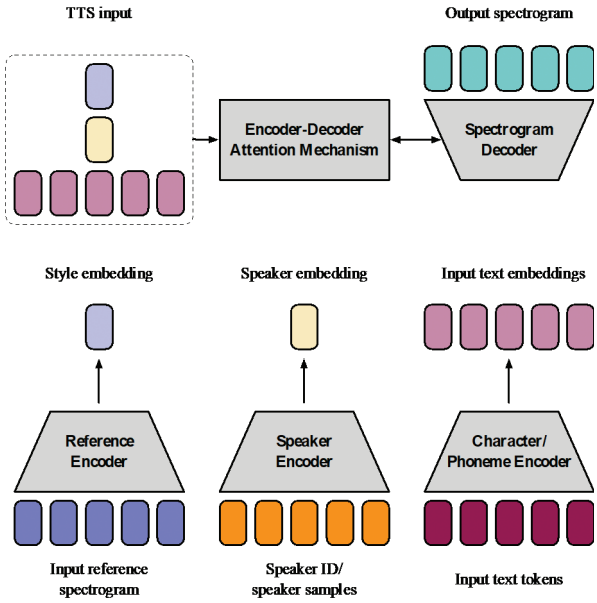
[2] In this context we define a *phoneme* as any of the perceptually distinct units of sound in a specified language that distinguishes one word from another, for example, p, b, d, and t in the English words pad, pat, bad, and bat.

encoder is used a reference from the decoder to align the generated speech to the input text to utter (see Fig. 1).

Between the encoder and decoder, there is an attention mechanism that allows learning implicitly the alignment between input and output. The activation of the attention can be retained and used to visualise this alignment.

The decoder usually generates the output autoregressively, consuming the latest generated slice of spectrogram (thus a spectrum) to output the following one. During training this approach allows guiding the generation process providing as input always the correct spectrum, taking it from the reference and speeding up the training process. At inference time, however, the decoder works completely on its own.

Figure 1 - *Example of encoder-decoder TTS synthesis with speaker and speaking style conditioning*



Tacotron (Shen, Pang, Weiss, Schuster, Jaitly, Yang, Chen, Zhang, Wang, Skerry-Ryan, Saurous, Agiomyrgiannakis, Wu, 2018; Wang, Skerry-Ryan, Stanton, Wu, Weiss, Jaitly, Yang, Xiao, Chen, Bengio, Le, Agiomyrgiannakis, Saurous Clark, Saurous, 2017) and DeepVoice (Ping, Peng, Gibiansky, Arik, Kannan, Narang, Raiman, Miller, 2018) are two examples of encoder-decoder models. The former uses recurrent neural networks as the main component while the latter uses convolutional networks. These two architectures have a separate module to predict the stopping point of the generation process.

FastSpeech (Ren, Ruan, Tan, Qin, Zhao, Zhao, Liu, 2019) represents instead a very different example of TTS. It uses a feed-forward architecture going from the input text to the output speech in a single module. The main architecture is based on that of an encoder-only transformer, processing in parallel the input and

generating the output. In this case, there are some additional modules to predict the phone duration. The encoder takes as input a sequence of graphemes or phonemes. Some models even allow using both encodings. The symbols in the input sequence are considered orthogonal symbols and are used to fetch embedding vectors that are processed by the encoder. The encoder usually leverages a bi-directional approach to find a latent representation of the input that can be used to condition the decoder to generate the correct spectrogram.

To complete the speech synthesis pipeline, a vocoder is necessary. The vocoder converts the output of the spectrogram predictor into an intelligible waveform. Neural vocoders have become a fundamental module of TTS models; they are necessary to synthesise a speech as clearly as possible (Jurafsky, Martin, 2022). Compared to the previous approach, the Griffin-Lim algorithm (Griffin, Lim, 1983), neural vocoders yield audio with fewer artefacts and higher quality.

Available implementations are based on (dilated) convolutional neural networks (e.g., WaveNet, Oord, Dieleman, Zen, Simonyan, Vinyals, Graves, Kalch-Brenner, Senior, Kavukcuoglu, 2016; WaveGlow, Prenger, Valle, Catanzaro, 2019; or MelGAN, Kumar *et al.*, 2019; Yang *et al.*, 2021) or recurrent neural networks (e.g., WaveRNN, Kalchbrenner, Elsen, Simonyan, Noury, Casagrande, Lockhart, Stimberg, Oord, Dieleman, Kavukcuogl., 2018).

There are two approaches to training these models. They can be either trained to work directly on raw waveforms (Oord *et al.*, 2016) or Mel-spectrograms (Yang *et al.*, 2021).

2.2 Expressive speech synthesis

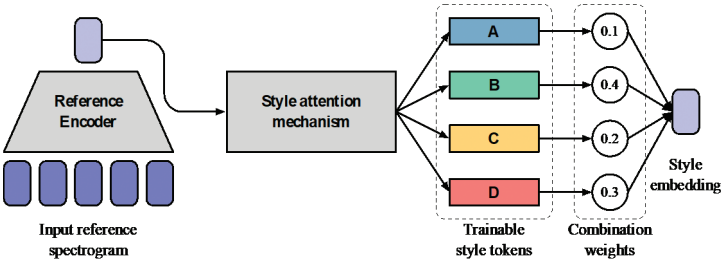
Besides the conditioning on the synthesised speech given by the input text (what to say), there are multiple research lines focused on controlling further aspects of the synthesised speech through additional information. These additional aspects can be categorised into two groups: speaker (or timbre, i.e., who is speaking) and prosody (or style, i.e., how to speak) (Tan *et al.*, 2021). These two aspects are completely orthogonal and can be combined (Skerry-Ryan, Battenberg, Xiao, Wang, Stanton, Shor, Weiss, Clark, Saurous, 2018).

Speaker control allows disentangling the uttered content from the speaker's voice, making the overall TTS model more reusable. The idea is to provide additional information on the speaker to the TTS. This approach allows leveraging multi-speaker data sets, while, previously, neural TTS used to be trained on single-speaker data sets (Wang *et al.*, 2017). To implement this kind of conditioning, it is sufficient to concatenate the hidden features extracted from a speaker recognition network to the hidden representation of the input text, as it was done for Tacotron 2 (Jia, Zhang, Weiss, Wang, Shen, Ren, Chen, Nguyen, Pang, Lopez-moreno, Wu., 2018; Favaro, Sbattella, Tedesco, Scotti., 2021; Favaro, Sbattella, Tedesco, Scotti, 2023) and DeepVoice 3 (Ping *et al.*, 2018).

In this paper we refer to the subtractive definition of prosody from Skerry-Ryan *et al.* (2018), where they consider prosody as the variation in speech signals

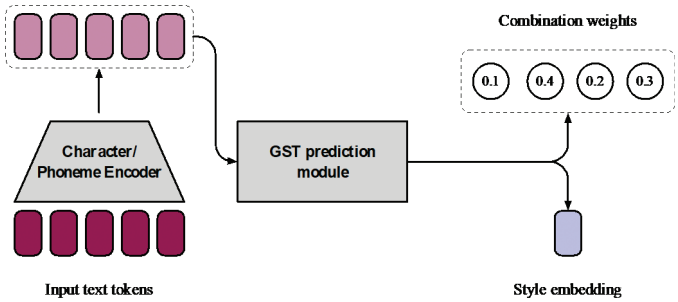
that remains after accounting for variation due to phonetics, speaker identity, and channel effects. In recent work on expressive TTS, prosody and style are used interchangeably. Prosodic control covers all the aspects concerning intonation, stress, and rhythm, which characterise how a sentence is uttered. These aspects also influence the emotion perceived by the listener. A reference step towards this kind of control on the speaking style is represented by the GSTs (Wang *et al.*, 2018). Instead of explicitly modelling the aspect characterising the prosody, this model uses unsupervised style representations learnt alongside the synthesis model. In the original implementation, a Tacotron 2 model was extended with a separate encoder that extracted a vector representing the so-called style embedding; this vector is concatenated to the hidden representation of the input text to provide the decoder with the style information (Wang *et al.*, 2018). We represented the GST computation process in Fig. 2. Notice that this kind of control works at a high level: the specific changes connected to a given style are learnt and implemented by the spectrogram generator.

Figure 2 - *Style embedding extraction using GSTs and combination weights from reference audio*



GSTs are designed to be learned from audio and used by extracting them from a reference clip. This would require having an available database of reference clips, which is not always a viable solution. Recently, approaches have been proposed to predict the GST from the raw text (Stanton, Wang, Skerry-Ryan, 2018) and solve this issue (see Fig. 3). There are two possible approaches: either predict the GST combination weights or directly predict the style embedding. The former is like a classification problem, where the output is a distribution over the components of the actual style embedding, the latter is instead a regression problem, where the style embedding is the target variable.

Figure 3 - *Style embedding and GST weights inference using from input text phonemes*



3. *Methodology*

In this section, we present DGST, explaining how it predicts the speaking style and how it is trained to predict such speaking style. This model revolves around the concept of style embedding, a (latent) vector representation of style in a dense continuous space.

3.1 Overview

In GST-powered speech synthesis modules, there is a sub-module in the network generating a latent vector representation of the speaking style, used to condition the speech synthesis (Valle *et al.*, 2020; Wang *et al.*, 2018). This latent vector is computed as the weighted combination of parameter vectors internal to the network. The weights of the combination are computed with the attention mechanism: a reference audio clip is processed into a single vector used as a query for attention, while the TTS provides the vectors to be used to compute keys and values. Some models, like Mellotron, adopt a multi-head attention approach, producing multiple output vectors (as many as the attention heads) and combining them (in the case of Mellotron, the vectors are simply concatenated).

Usually, the style is extracted from a reference speech clip, which may not be available at run time. Previous solutions for GST prediction from text proposed by Stanton *et al.* (2018), focused on predicting the style simply from the text to utter, predicting either the prosody embedding directly or predicting the combination weights to be passed to the attention of the style encoder in the TTS. They used as input the encoding of the text to utter realised by the TTS, which is usually computed from the phonemes.

The module we build, DGST, proposes to focus on the dialogue domain, also including information from the context of the conversation. We leverage the hidden representations of a language model, which are usually very informative and yield impressive results on many regression and classification tasks.

3.2 DGST model

The model we propose combines two tasks together: a regression task to predict the style embedding and a classification task to predict the style combination weights (the combination weights are a distribution of probability over the available style tokens, i.e., the latent codes).

DGST takes as input a 1D feature map (i.e., a matrix) $H \in R^{t \times d}$ (t is the number of elements in the input sequence and d is the number of dimensions of the vectors encoding a single sequence element) encoding the text the TTS needs to utter. Each column vector in this feature map is an embedding computed by a deep learning-based language model.

DGST learns two output functions $f_{GSTe}(\boldsymbol{H}) = \hat{\boldsymbol{y}}_{GSTe} \in \mathbb{R}^{d_e}$ (d_e is the number of dimensions of a style embedding) and $f_{GSTw}(\boldsymbol{H}) = \hat{\boldsymbol{Y}}_{GSTw} \in \mathbb{R}^{h \times d_h}$ (h is the number of attention heads and d_h is the number of dimensions of a single attention head), both use as input the 1D feature map. The former tackles the problem of predicting directly the style embedding as if it was a regression problem. The latter tackles the problem of predicting the combination weights of style embeddings (note that we are considering a generic case where the GST module uses h different heads for the combinations weights on the style tokens).

What characterises DGST is the choice of how the input features are computed. We considered three possible approaches:

1. The input embeddings are computed from the current turn to utter in the dialogue and nothing else, as in Fig. 4. The model ignores the dialogue context and uses only information from the current response.
2. The input embeddings are computed from the current turn to utter in the dialogue and the preceding turn, but only the embeddings of the current turn are used by DGST, as in Fig. 5. Since deep learning-based language model exploit context to compute the hidden representations, even if we use only the embeddings from the last turn, they include information coming from the context.
3. The input embeddings are computed from the current turn to utter in the dialogue and the preceding turn, and all the embeddings are used by DGST, as in Fig. 6. In this case we are adding explicitly the information from the embeddings of previous turns.

DGST works entirely with latent representations, from input to output. It predicts the most probable style embedding or embedding combination weights from the contextual latent representation yield by a language model. Given these latent space interconnections, DGST is bound to the language model and expressive TTS used to compute its inputs during training. These modules cannot be changed or DGST would have to be re-trained.

Figure 4 - *Style embedding and GST weights inference from dialogue response contextual embeddings*

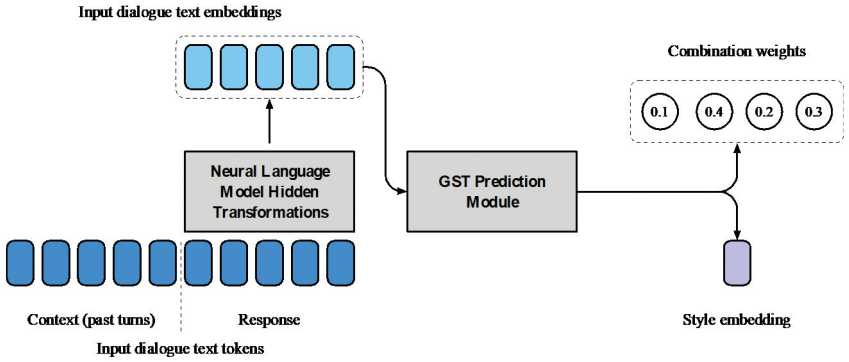


Figure 5 - *Style embedding and GST weights inference from dialogue response contextual embeddings (computed using also dialogue context)*

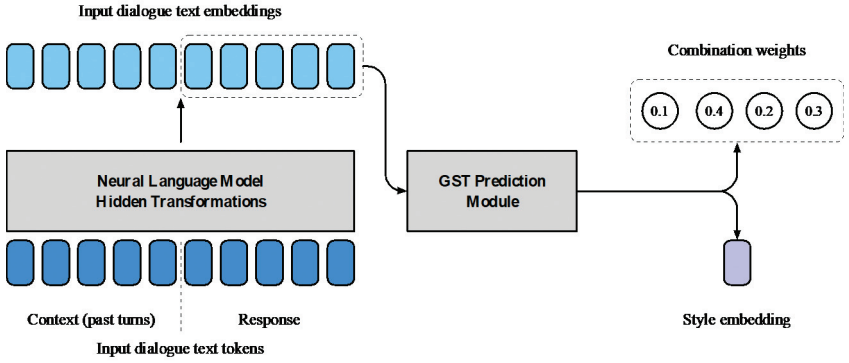
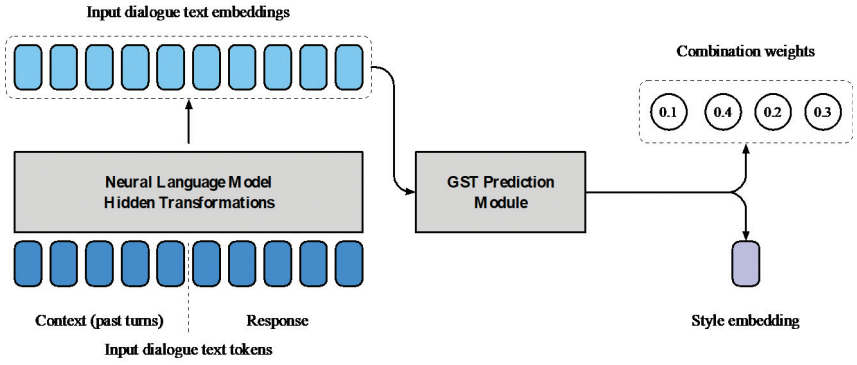


Figure 6 - *Style embedding and GST weights inference from dialogue context and response contextual embeddings*



3.3 Training approach

To train this network we employ two external models: a dialogue language model and a TTS with GST conditioning. The Dialogue language model is used to extract the contextual embeddings H that feed our model (i.e., the input), while the GST module of the TTS is used to extract the targets $\boldsymbol{y}_{GSTe} \in \mathbb{R}^{d_e}$ and $\boldsymbol{Y}_{GSTw} \in \mathbb{R}^{h \times d_h}$ from audio clips containing the speech uttering the response processed by the dialogue language model.

The overall loss function $\mathcal{L}_{DGST}(\cdot)$ is described in (1).

$$(1)\ \mathcal{L}_{DGST}(\boldsymbol{H}, \boldsymbol{y}_{GSTe}, \boldsymbol{Y}_{GSTw}) = \mathcal{L}_{DGSTe}(\boldsymbol{H}, \boldsymbol{y}_{GSTe}) + \mathcal{L}_{DGSTw}(\boldsymbol{H}, \boldsymbol{Y}_{GSTw})$$

It is the sum of two losses, one for the embedding prediction head $\mathcal{L}_{DGSTe}(\cdot)$ and one for the combination weights prediction head $\mathcal{L}_{DGSTw}(\cdot)$. In the reference work by Stanton *et al.* (2018), they used the L1 norm loss for both heads, here instead, we distinguish between the two tasks of the heads.

Given that the embedding prediction head behaves as a regression model (i.e., a model predicting a continuously valued output, which is the embedding in our case), we use the mean squared error between the embedding prediction and the target embedding as loss function, as described by (2).

$$(2)\ \mathcal{L}_{DGSTe}(\boldsymbol{H}, \boldsymbol{y}_{GSTe}) = \frac{1}{d_e} \| \boldsymbol{y}_{GSTe} - f_{DGSTe}(\boldsymbol{H}) \|_2^2$$

Since the combination weights can be seen as a discrete distribution, we compute the loss of the combination weights prediction head as the distance between the reference and target distributions using the Kullback–Leibler (KL) Divergence. The KL Divergence $D_{KL}\,[P \parallel Q]$ of P from Q is a measure of how the probability distribution P is different from a reference probability distribution Q. We computed the average KL Divergence between the target weights and the predicted combination weights for each head. The KL Divergence is computed head-wise, and then we averaged the score among the heads as in (3).

$$(3)\ \mathcal{L}_{DGSTw}(\boldsymbol{H}, \boldsymbol{Y}_{GSTe}) = \frac{1}{h} \sum_{i=1}^{h} D_{KL}[\boldsymbol{Y}_{GSTw,i} \parallel f_{DGSTw}(\boldsymbol{H})_i]$$

4. *Experiments and results*

In this section we present the training and evaluation approaches we adopted with DGST (including the details on the experiments configurations). Subsequently we analyse and comment the obtained results.

4.1 Experimental approach

We trained multiple versions of DGST varying the underlying language model to extract the input features and varying how to extract such features. In any case, we trained for the prediction of both GST embedding and combination weights using a single model, exploiting the same hidden representation.

Concerning the language models, to observe the effects of their complexity and size, we evaluated three versions of GPT-2 fine-tuned on large dialogue corpora. To this end, we resorted to DialoGPT (Zhang, Sun, Galley, Chen, Brockett, Gao, Gao, Liu, Dolan, 2020), using the 117M, 345M and 762M parameter versions. Additionally, we repeated the same training steps using Therapy-DLDLM (Scotti, 2023), a custom fine-tuning of GPT-2 (762M parameter version) for dialogue that uses a discrete latent representation to encode dialogue turn. We choose this additional model to see whether exploiting its feature vectors (that can compress entire dialogue turns) would reach the same or comparable results of a more generic model pre-trained on larger data sets (Therapy-DLDLM should be the language model we are interested in using for the empathetic chatbot we are willing to provide with expressive speech capabilities).

We trained different models adopting all the feature extraction approaches presented in § 3. We used all the language models' hidden transformations to extract the sequence of embeddings from the response only, from the context and the response keeping only the hidden vectors of the response, and from the context and the response keeping all hidden vectors.

The actual DGST model is a single layer of Gated Recurrent Units (GRUs, Cho, Van Merrienboer, Gülçehre, C, Bahdanau, Bougares, Schwenk, Bengio., 2014) to process the input sequence followed by two alternative linear projections (one per target). Before the GRU layer, we applied dropout with a 10 % probability to regularise the model. We found that this configuration was sufficient to obtain good results.

We trained all models with a constant learning rate of $\eta = 5 \cdot 10^{-4}$ and a batch size of 32 elements. During training, we found that within 5 epochs, all models would reach a plateau on the validation scores. We used AdamW (Loshchilov, Hutter, 2019) as optimiser.

The main evaluation we conducted on this model is based on the training metrics. We compare the final test values, considering all the model variants we considered to understand the role of model complexity and dialogue context when predicting the GST embedding or the GST combination weights from the text.

To have statistical evidence that the results of the best performing model variant were the best, we run a t-test on these results. We compared the scores achieved by the best model with those of all the other variants, computing the p-value for statistical significance with the confidence level $\alpha = 0.05$. The null hypothesis was that the two populations had same mean and with a p-value $< \alpha$ we could reject that hypothesis. We run the test separately on KL-Divergence scores and MSE scores, in this way we could compare separately the combination weights prediction and the embedding prediction.

Additionally, we performed an evaluation using the spectrogram generator. We compared the Mel-spectrogram synthesised by the TTS using the reference clip and the GST embedding or the GST combination weights predicted by DGST using the best-performing configuration.

4.2 Data

We conducted the training and evaluation session of DGST on the IEMOCAP corpus (Busso, Bulut, Lee, Kazemzadeh, Mower, Ki Chang, Lee, Narayanan, 2008). IEMOCAP is a multimodal corpus for dialogue. This data set offers scripted and spontaneous dialogues in English displaying different emotions that result in a wide variety of speaking styles.

Our objective is to exploit this rich variety of speaking style characterising the corpus. The neural network implementing the DGST module should learn to choose the appropriate speaking style given the text to utter (i.e., the latest response in a dialogue) and, possibly, the conversational context (i.e., conversation history). Thanks to the audio recordings, we can use a pre-trained speaking style encoder coming from a TTS model to compute the targets to predict from input. In this work we used the style encoder of a Mellotron TTS model[3].

Table 1 - *IEMOCAP data set main statistics*

	Training	*Validation*	*Test*
Dialogues	84	29	38
Utterances	5754	1818	2373

In Tab. 1 we reported the main statistics on the corpus. Differently from common benchmark for chatbots, IEMOCAP does not contain many dialogues; however, the task is simpler than learning how to respond in a dialogue. Moreover, the individual conversations are longer. This should help the model learn long term dependencies.

4.3 Results and comments

We report the results on the target metrics for the two DGST approaches in Fig. 7 and Fig. 8. The scores show interesting behaviour.

From the results on DialoGPT it seems that the size of the language model leads to overfitting and that the 345M parameters version reaches the best performances. Moreover, the results seem to indicate that using the information from the context negatively affects the performances with respect to processing only the response. However, the results obtained using the features computed by Therapy-DLDLM highlight a completely different result.

We achieved the best results using the most complex language model (Therapy-DLDLM is built from a 762M parameters GPT-2) and processing with DGST the embeddings of the response computed also using the context. We hypothesise that the hierarchical nature of the variational model in Therapy-DLDLM enforces the

[3] We used a custom wrapper (link: https://github.com/vincenzo-scotti/tts_mellotron_api) over the original Mellotron TTS model (link: https://github.com/NVIDIA/mellotron); the models we used takes as input a mixture of phonemes (when the word to utter belongs to a vocabulary of pre-encoded words) and graphemes (when the word to utter is out of the reference vocabulary).

latent representation to extract a contextual embedding better summarising the sequence. Another possible explanation is in the quality of data, since DialoGPT was trained on a larger data set that was not manually curated as those used for Therapy-DLDLM; so, possibly, to learn a good model from these data sets a larger number of samples may be required.

Figure 7 - *Test set MSE on style embedding reconstruction using different approaches for input embeddings and different language models*

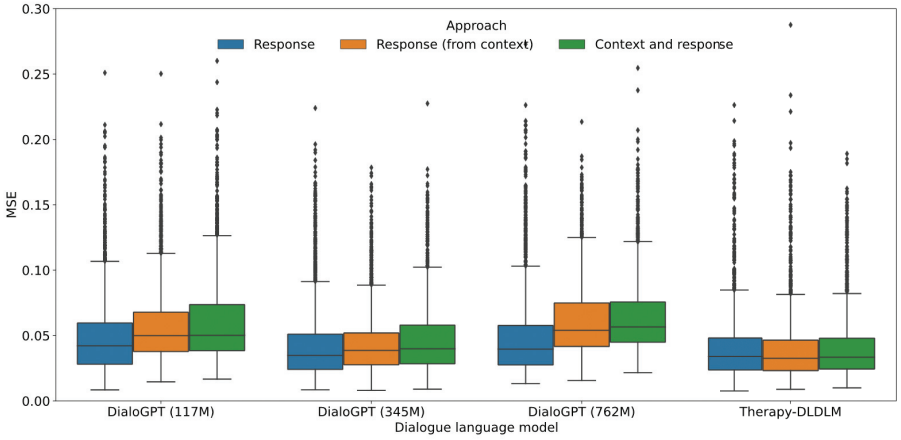


Figure 8 - *Test set KL-Divergence on GST weights distribution reconstruction using different approaches for input embeddings and different language models*

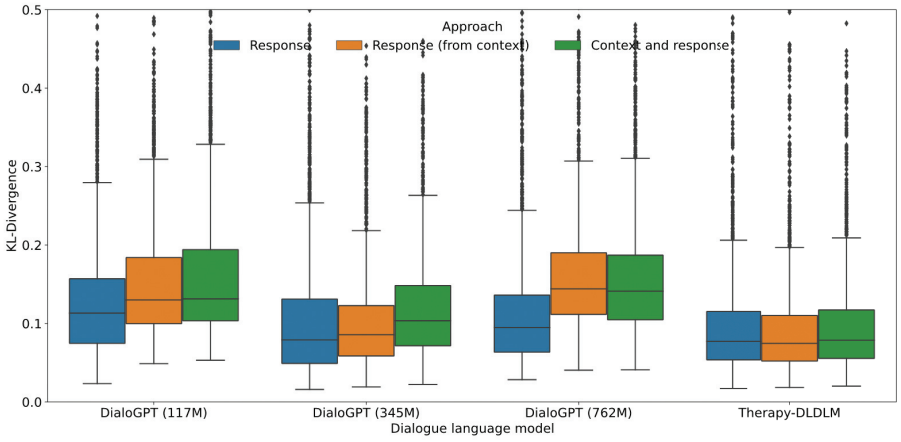


Table 2 - *T-test results comparing the best performing DGST model (DGST with Therapy-DLDLM contextual embeddings of dialogue response, computed also using the dialogue context) with all the other DGST versions*

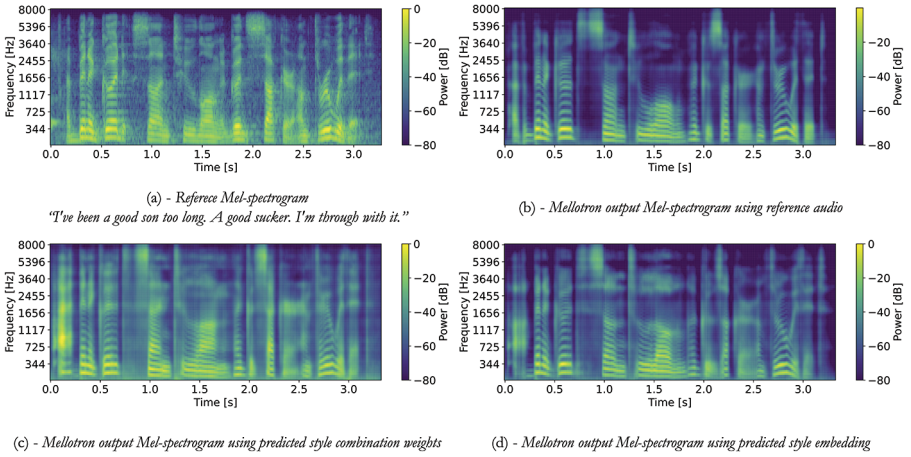
Approach	*p-value*			
	DialoGPT (117M)	*DialoGPT (345M)*	*DialoGPT (762M)*	*Therapy-DLDLM*
	MSE			
Response	1.777×10^{-34}	2.941×10^{-7}	4.707×10^{-33}	0.076
Response (from context)	5.310×10^{-114}	7.236×10^{-10}	3.015×10^{-172}	-
Context and response	1.635×10^{-136}	8.380×10^{-25}	4.181×10^{-203}	0.150
	KL-Divergence			
Response	1.218×10^{-74}	1.920×10^{-13}	1.202×10^{-29}	0.012
Response (from context)	1.361×10^{-180}	1.466×10^{-8}	3.930×10^{-233}	-
Context and response	9.014×10^{-193}	7.022×10^{-49}	4.401×10^{-208}	0.005

In Tab. 2 we report the p-values from the t-tests comparing the DGST model using Therapy-DLDLM transformer with the "response from context" encoding approach, which was the best-performing model on both approaches. In all tests, we set $\alpha = 0.05$ and, thus, a confidence level of 95 %. If compared to the models based on DialoGPT, DGST based on Therapy-DLDLM using the "response from context" not only performs better than the other models but has evidence that the difference in the results is statistically significant. When we compare the results of DGST with the same Therapy-DLDLM and the different encoding approaches, we can see that the results are statistically significant only on the KL-Divergence score, and thus the models predicting the GST combination weights. Differently, on the MSE scores (the models directly predicting the GST embeddings), despite we get better results, we don't have the same statistical evidence supporting the difference with the other encoding approaches.

To get an idea of the actual effect of the predicted GSTs on the synthesis, we visualised the generated Mel-spectrograms. We report an example in Fig. 9 comparing the original spectrogram with the one generated by Mellotron using the original audio clip for the style input and the two generated spectrograms using Mellotron and the GST predicted using the two approaches proposed with DGST.

The prediction using the GST embeddings has a noisier spectrogram, with an overall higher average power, as a result, the image of the Mel-spectrogram appears brighter and more like that of the source audio clip. The prediction using the GST combination weights is less noisy and is closer to the one generated using the original clip. In both cases, the spectrogram does not appear distorted and is not much different from the one generated using the two audio clips (look for example at the harmonics). This similarity indicates that both prediction approaches (either GST embeddings or GST combination weights) may be equally valid.

Figure 9 - *Comparison of reference Mel-spectrogram with synthesised Mel-spectrogram using Mellotron*



(a) - *Reference Mel-spectrogram*
"I've been a good son too long. A good sucker. I'm through with it."

(b) - *Mellotron output Mel-spectrogram using reference audio*

(c) - *Mellotron output Mel-spectrogram using predicted style combination weights*

(d) - *Mellotron output Mel-spectrogram using predicted style embedding*

5. *Conclusion*

In this paper we present a methodology to train a model to predict speaking style from text: DGST. DGST showed to be a valid tool to bridge text and speech, allowing control of the expressiveness of the speech by learning from human examples during conversations. The interesting finding is in the input features and the approach to extracting such features.

The best performances in terms of GST prediction, either considering GST embeddings or considering GST combination weights, are reached using the contextual embeddings yield by Therapy-DLDLM. However, we got statistical evidence on the results only when comparing models predicting the GST combination weights and not the model predicting the GST embeddings. Moreover, Therapy-DLDLM compares better than the same size and same architecture models trained on larger data sets (namely DialoGPT). We hypothesise this is due to the hierarchical nature of the variational model. Due to how Therapy-DLDLM processes the dialogue, the hidden representation of text computed by the model is forced to summarise the entire sequence, helping the computation of GSTs.

Including dialogue context seems to help the recognition of the GST to use for the response. However, rather than processing with the DGST module both context and response, processing only the embeddings of the response, encoded also using the context information, yields empirically the best results. We suppose that this is a consequence of using GRU for sequence modelling, which despite being a powerful sequence modelling tool, as other recurrent networks fails when the context gets too large. Nevertheless, using GRU allowed impressive results without resorting to the fine-tuning of the transformer dialogue model, which requires more computational effort and may need more samples to be correctly fine-tuned.

The next step we are considering for this kind of modules to connect speech and text is to drop the latent representation and start using explicit names to address the desired speaking style. Similarly to current solutions for image generation from text, where it is possible to ask for the desired style of the image (e.g., photorealistic image or oil painting), we are interested in asking through a textual prompt for the speaking style (e.g., loudly or mumbling). A solution of this kind would make the model easier to interpret and use.

Appendix

For transparency and reproducibility, we release the developed source code via GitHub at the following link: https://github.com/vincenzo-scotti/dialogue_gst

References

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H., (a cura di), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 6-12 December 2020, virtual.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim. S., Chang, J.N., Lee, S., and Narayanan, S.S., (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4): 335–359.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Moschitti, A., Pang, B., Daelemans, *W., Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, 25-29 October 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 1724–1734. ACL. doi: 10.3115/V1/D14-1179.

Google, 2023, Introducing BARD. URL: https://blog.google/technology/ai/bard-google-ai-search-updates/

Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez-Moreno, I., Wu, Y., (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. A cura di Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi,N., Garnett, R., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, Montréal, Canada, 3-8 December 2018, 4485–4495.

Jurafsky, D., Martin, J.H., (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, 2nd Edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Favaro, A., Sbattella, L., Tedesco, R., and Scotti, V., (2021). Itacotron 2: Transfering english speech synthesis architectures and speech features to italian. In Abbas, M., Alhakim Freihat, A., (a cura di), *4th International Conference on Natural Language and Speech Processing*, Trento, Italy, 12-13 November 2021, 81–86. URL https://aclanthology.org/2021.icnlsp-1.10

Favaro, A., Sbattella, L., Tedesco, R., & Scotti, V., (2023). ITAcotron 2: The Power of Transfer Learning in Expressive TTS Synthesis, Springer International Publishing, Cham, 1-20. doi: 10.1007/978-3-031-11035-1_1.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, V.D., A., Dieleman, S., Kavukcuogl, K., (2018). Efficient neural audio synthesis. In. Dy, J.G., Krause, A., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Stockholmsmässan, Stockholm, Sweden, 10-15 July 2018, vol. 80 of Proceedings of Machine Learning Research, 2415–2424.

Loshchilov, I., Hutter, F., (2019). Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019,* New Orleans, LA, USA, 6-9 May 2019.

OpenAI, 2022, Introducing ChatGPT, URL: https://openai.com/blog/chatgpt

Paiva, A., Leite, I., Boukricha, H., Wachsmuth, I., (2017). Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.*, 7(3):11:1–11:4. doi: 10.1145/2912150.

Picard, W.P., (1997). *Affective computing*. MIT press.

Ping, W., Peng, K., Gibiansky, A., Arik, S.Ö., Kannan, A., Narang, S., Raiman, J., Miller, J., (2018). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, 30 April – 3 May 2018, Conference Track Proceedings. OpenReview.net.

Prenger, R., Valle, R., Catanzaro, B., 2019, Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, Brighton, United Kingdom, 12-17 May 12019, 3617–3621.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., (2022). Robust Speech Recognition via Large-Scale Weak Supervision. In: *CoRR* abs/2212.04356. DOI: 10.48550/ARXIV.2212.04356.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. In *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T., (2019). Fastspeech: Fast, robust and controllable text to speech. In Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., (a cura di), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Vancouver, BC, Canada 8-14 December 2019, 3165–3174.

Scotti, V., (2023). *Generative Empathetic Data-Driven Conversational Agents for Mental Healthcare.* PhD thesis. Polytechnic University of Milan, Italy. Handle: https://hdl.handle.net/10589/202713

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y.,

(2018). Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, Calgary, AB, Canada, 15-20 April 2018, 4779–4783.

Skerry-Ryan, R.J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R.J., Clark, R., Saurous, R. A, (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In Dy, J., G., Krause, A., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018,* Stockholmsmässan, Stockholm, Sweden, 10-15 July 2018, vol.80 of Proceedings of Machine Learning Research, 4700–4709.

Stanton, D., Wang, Y., Skerry-Ryan, R.J., (2021). Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018*, Athens, Greece, 18-21 December 2018, 595–602.

Tan, X, Qin, T., Soong, F.K., Liu, T.Y., (2021). *A survey on neural speech synthesis*. CoRR, abs/2106.15561.

Valle, R., Li, J., Prenger, R., Catanzaro, B., (2020). Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020*, Barcelona, Spain, 4-8 May 4-8 2020, 6189–6193.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K., (2016). Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, 13-15 September 2016, 125.

Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Saurous Clark, R., Saurous, R.A., (2017). Tacotron: Towards end-to-end speech synthesis. In Lacerda, F., *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 20-24 August 2017, 4006–4010.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A., (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Dy, J.G., Krause, A., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Stockholmsmässan, Stockholm, Sweden, 10-15 July 2018, vol. 80 of Proceedings of Machine Learning Research, 5167–5176.

Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., Gao, X., Gao, J., Liu, J., Dolan, B., (2020). DIALOGPT: Large-scale generative pre-training for conversational response generation. In Celikyilmaz, A., Wen, T.-H., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020*, Online, 5-10 July 5-10 2020, 270–278.

DOI: 10.17469/O2111AISV000021

FEDERICA CAVICCHIO, MIRKO GRIMALDI

I training audio e audiovisivi migliorano in misura uguale la percezione segmentale della seconda lingua

Audio and audiovisual training equally enhance second language segmental perception

This study investigates the efficacy of high variability phonetic training, using audiovisual stimuli, in improving the perception of L2 Southern Standard British English (SSBE) phonemes by L1 Italian speakers. Participants underwent an audio pre-test and were assigned to audio, audiovisual, or control training groups. The post-test results revealed that audio and audiovisual training significantly improved the phoneme identification accuracy compared to the control group. However, during the training, the audiovisual information positively influenced the identification of L2 vowels. This improvement is, however, temporary. According to the *Analysis by Synthesis model*, our findings suggest that learners have the acoustic signal as privileged cue for building their phonetic-phonological representations. Their learning is facilitated only when the acoustic signal contains information related to distinctive features already active in the contrastive system of the L1. Further exploration of the interactions between auditory and articulatory information is essential for developing effective teaching approaches to L2 phonemes perception.

Keywords: high variability phonetic training, audiovisual speech, second language learning, second language sound perception.

1. *Introduzione*

L'*High Variability Phonetic Training* (HVPT) è uno dei metodi più utilizzati per migliorare le abilità di percepire e produrre i suoni di una seconda lingua straniera (L2). Sebbene il termine *High Variability Phonetic Training* sia stato coniato da Iverson, Hazan & Bannister (2005), le prime ricerche che utilizzano l'HVPT sono iniziate ben prima, stimolate dal dibattito sull'ipotesi del periodo critico di acquisizione della lingua e la sua validità in merito all'acquisizione della L2 (Flege, 1987; Major, 1987). Con l'accumularsi nel tempo di ricerche che usano l'HVPT, questo metodo si è confermato in grado di generare cambiamenti misurabili nella percezione e produzione dei fonemi L2 degli studenti adulti.

Nel primo studio in cui l'HVPT è stato utilizzato, Logan, Lively & Pisoni (1991) hanno addestrato sei apprendenti giapponesi a identificare il contrasto inglese /l/-/r/, assente nel loro sistema consonantico, usando un compito di identificazione seguito da feedback. Gli studenti sono stati esposti a 68 coppie minime di parole contenenti i suoni consonantici /l/-/r/in posizioni iniziale, mediale o finale di sillaba

pronunciate da cinque diversi parlanti. Dopo 40 minuti di training, gli studenti erano significativamente più abili a discriminare /l/-/r/e questa conoscenza si era generalizzata sia rispetto a parole nuove non addestrate sia rispetto ad un nuovo parlante che gli studenti non avevano sentito durante l'addestramento. L'accuratezza variava in base al contesto fonetico in cui si verificava il contrasto /l/-/r/. In particolare, gli studenti erano più bravi a identificare il contrasto quando si trovava in posizione finale di sillaba rispetto alla posizione iniziale. Logan *et al.* (1991) hanno anche scoperto che il contesto influenzava il tempo di reazione. Studi successivi hanno confermato i risultati di Logan *et al.*, 1991, generalizzandoli anche rispetto ai fonemi vocalici. Per quanto riguarda gli stimoli uditivi, nella maggior parte degli studi questi erano prodotti da 2 fino a 30 parlanti nativi. Mentre la norma è usare stimoli naturali per l'addestramento e i test di verifica, uno studio ha anche utilizzato la sintesi vocale per produrre gli stimoli L2 (Qian, Chukharev-Hudilainen & Levis, 2018).

In ricerche sulla comprensione lessicale, svariati studi hanno impiegato metodologie come il priming (per esempio, Broersma, 2012) e l'eye-tracking (ad esempio, Cutler, Weber & Otake, 2006; Escudero, Hayes-Herb & Mitterer, 2008; Weber e Cutler, 2004) per analizzare le differenze nell'elaborazione linguistica e lessicale tra parlanti L2 e parlanti nativi (L1). I risultati indicano che i parlanti L2, a differenza dei parlanti L1, tendono a mostrare tempi di reazione più lunghi per le parole che presentano contrasti fonetici tipici della L2, come ad esempio /æ/-/ʌ/. Invece, i parlanti L1 reagiscono più rapidamente a tali contrasti. Questo suggerisce che esiste una differenza significativa nel modo in cui i parlanti nativi e quelli L2 attivano lessicalmente le parole dei contrasti fonetici caratteristici di una determinata lingua. Il contrasto viene quindi percettivamente assimilato a un suono L1 (ad esempio, la vocale L1 /a/ per il contrasto /kæp/-/kʌp/). La difficoltà nel distinguere suoni L2 percettivamente simili aumenta la competizione lessicale, rendendo difficile il riconoscimento delle parole e potenzialmente portando alla confusione delle parole L2 in coppia minima. Neppure un'accurata categorizzazione dei suoni contrastivi in L2 sembra garantire un'accurata discriminazione a livello lessicale; infatti, le parole in coppia minima possono essere percepite come omofone se il contrasto target non è stato accuratamente codificato lessicalmente dai parlanti L2 (Darcy, Dekydtspotter, Sprouse, *et al.*, 2012). Di conseguenza, l'addestramento HVPT con parole in coppia minima può essere meno efficace rispetto all'uso di non-parole per le quali i parlanti L2 non hanno ancora sviluppato rappresentazioni lessicali. Thomson (2011) ha condotto una serie di training con HVPT modificando i contesti degli stimoli uditivi. Un gruppo di studenti L2 è stato addestrato a percepire vocali inglesi all'interno di non-parole CV prodotte da parlanti nativi, mentre un altro gruppo è stato addestrato con non-parole manipolate sinteticamente in modo tale che le vocali fossero raddoppiate in lunghezza. Entrambi i gruppi hanno dimostrato un miglioramento nell'identificazione delle vocali addestrate a prescindere dal tratto di lunghezza delle vocali.

1.1 Stimoli e tipi di training con l'HVPT

Nel HVPT gli stimoli uditivi vengono solitamente presentati in sessioni bloccate per parlante L1, come suggerito da Logan *et al.* (1991) e Lively, Logan & Pisoni (1993), invece di utilizzare più parlanti all'interno di una singola seduta di training. Tuttavia, studi più recenti hanno usato fino a 30 parlanti L1 per sessione (ad esempio, Thomson, 2016a; Thomson, Derwing, 2016).

L'uso del paradigma Forced Choice IDentification (FCID) sembra, in generale, essere più efficace della discriminazione AX, sia per le consonanti (Flege, 1995a) che per le vocali (Rato, Rauber, 2015; Carlet, 2017). Nel paradigma FCID, gli studenti devono identificare il fonema appena ascoltato, utilizzando simboli fonetici o ortografici (ad esempio, "hai sentito /l/ o /r/?" vs "hai sentito l o r?") dopo aver ascoltato uno stimolo uditivo. Resta invece ancora aperta la questione del training adattivo, ovvero dei cambiamenti della struttura del training in risposta alle prestazioni dei singoli studenti L2. Diversi studi hanno utilizzato una qualche forma di training HVPT adattivo (Pruit, Jenkings & Strange, 2006; Iverson, Evans, 2009; Lengeris, Hazan, 2010), ma nessuno di questi ha paragonato l'effetto dell'apprendimento adattivo con quello non adattivo.

La durata del training solitamente varia da poche ore ad alcune settimane, con risultati omogenei (Logan *et al.*, 1991; Thomson, 2016b). Il numero di sessioni di formazione e la loro durata cumulativa variano notevolmente da studio a studio. Ad esempio, Cebrian, Carlet (2014) hanno completato quattro sessioni di training, per un totale di tre ore, mentre i partecipanti di Nishi, Kewley-Port (2007, 2008) hanno completato nove sessioni, per un totale di 13,5 ore. Non sembra esistere una corrispondenza tra la durata del training e il numero di fonemi utilizzati per il training. Ad esempio, Logan *et al.* (1991) ha mirato solo al contrasto /l/-/r/ e Thomson (2016a) ha sviluppato un training per dieci vocali: in entrambi gli studi gli apprendenti sono stati sottoposti a dieci ore di training in totale senza un evidente vantaggio della durata dei training sul livello di apprendimento dei fonemi.

1.2 Generalizzazione e mantenimento nel tempo dell'apprendimento percettivo

Uno dei vantaggi più importanti dell'HVPT è che l'apprendimento percettivo dei suoni L2 oggetto del training si generalizza rispetto a nuove istanze dello stesso suono e l'apprendimento viene mantenuto nel tempo. L'effetto sull'apprendimento percettivo L2 è solitamente generalizzato anche rispetto a nuovi parlanti che hanno prodotto le stesse non-parole o sillabe usate durante il training (Lively *et al.*, 1993). Tuttavia, la generalizzazione potrebbe non avvenire per tutti i suoni oggetto del training, come ad esempio nell'inglese /ð/ (Cebrian, Carlet, 2014).

Ci sono risultati contrastanti riguardo ai test di generalizzazione dell'HVPT rispetto a nuove non-parole. Alcuni studi hanno mostrato che il training con parlanti multipli ha portato a discriminare anche nuove non-parole prodotte da nuovi parlanti (Lively *et al.*, 1993), mentre il training con un solo parlante non sembra portare alla generalizzazione di non-parole non familiari prodotte dallo stesso parlante. Altri studi hanno trovato solo un trasferimento parziale dell'apprendimento percettivo agli stessi suoni in nuovi

contesti. Iverson *et al.* (2005) hanno riportato che la percezione del contrasto inglese /l/ e /r/ da parte di parlanti L2 giapponese si estende a nuove parole, principalmente a quelle che iniziano con /l/ e /r/. L'apprendimento non era invece generalizzato quando le stesse consonanti si trovano in posizione mediale tra vocali. Thomson (2011) ha evidenziato che, in produzione, 10 vocali inglesi nei contesti /bV/ e /pV/ venivano generalizzate, dopo il training HVPT, anche in contesti /zV/ e /sV/, ma non in contesti /gV/ e /kV/. Tuttavia, in percezione è stato riscontrato il comportamento opposto. Le vocali del training in contesti /bV/ e /pV/ erano generalizzate nei contesti /gV/ e /kV/, ma non nei contesti /zV/ e /sV/ (Thomson, 2012). Altre ricerche invece non hanno trovato alcun trasferimento a nuovi contesti. Tajima *et al.* (2008) hanno riportato che, dopo il training, i parlanti anglofoni sono in grado di percepire la lunghezza dei fonemi del giapponese e di generalizzarla a nuovi parlanti, ma non hanno trovato un trasferimento della percezione vocalica a nuovi contesti.

Un'altra domanda rilevante nell'utilizzo dell'HVPT è fino a che punto sia possibile trasferire il training fonetico fatto in contesto di parole isolate a parole che si trovano nel contesto di un discorso. Hirata (2004) ha addestrato studenti inglesi a percepire i contrasti di lunghezza delle vocali e delle consonanti giapponesi. Un gruppo ha ascoltato parole pronunciate in isolamento, mentre un altro gruppo ha ascoltato le stesse parole pronunciate in una frase. I risultati dimostrano che entrambe le condizioni hanno portato a un significativo miglioramento. Inoltre, il trasferimento dall'apprendimento in contesti di frasi a quello in parole isolate era maggiore del trasferimento dall'addestramento di parole isolate al contesto in frasi.

L'apprendimento fonetico dovrebbe non solo essere generalizzabile, ma anche mantenuto nel lungo termine. Diversi studi hanno dimostrato come, dopo il training con HVPT, l'apprendimento fonetico è mantenuto dopo due settimane (Flege, 1995b), un mese (Thomson, 2012), due mesi (Rato, Rauber, 2015), tre mesi (Bradlow, Pisoni, Akahane-Yamada & Tohkura, 1997; Wang, Munro, 2004; Nishi, Kewley-Port, 2007; Carlet, 2017) e quattro mesi (Iverson, Evans, 2009). Solo due studi hanno indagato la persistenza dell'apprendimento oltre i quattro mesi. Lively, Pisoni, Yamada *et al.*, (1994) hanno scoperto che gli studenti giapponesi miglioravano la loro pronuncia del contrasto inglese /l/-/r/ fino a tre mesi dopo il training, ma dopo sei mesi si presentava un declino e i punteggi di identificazione non erano significativamente diversi dai punteggi pre-test. Viceversa, Wang, Spence, Jongman & Sereno (1999) hanno trovato che l'apprendimento percettivo può persistere senza declino per almeno sei mesi.

Thomson, Derwing (2016) hanno studiato se il training con non-parole portasse a un miglioramento nella produzione di 10 vocali inglesi L2 rispetto all'addestramento con parole L2 contenenti le medesime vocali. I 31 partecipanti (11 L1 cinesi, 5 L1 spagnoli, 3 L1 etiopi parlanti amarico, 3 L1 arabi, 2 L1 russi, 2 L1 francesi, 2 L1 Punjabi, 1 L1 rumeno, 1 L1 zo'é e 1 L1 igbo) sono stati divisi in due gruppi. Il primo gruppo è stato addestrato a percepire le dieci vocali inglesi con non-parole contenenti le vocali in contesto CVC. Questo gruppo ha significativamente migliorato la pronuncia di 70 parole reali contenenti le stesse vocali addestrate. Al contrario, il secondo gruppo addestrato a percepire le stesse vocali, utilizzando però le 70 parole reali, non aveva

migliorato significativamente la pronuncia delle parole stesse. Questo risultato è coerente con quanto riportato da Nishi, Kewley-Port (2007) e Carlet (2017), che nei loro studi hanno usato non-parole per addestrare la percezione fonetica delle vocali dell'inglese L2. Entrambi gli studi riportano che l'addestramento con non-parole veniva generalizzato a parole che contenevano le vocali addestrate.

1.3 HVPT con stimoli audiovisivi

È noto che, negli adulti, le aree senso-motorie e uditive sono attivate congiuntamente durante l'acquisizione dei suoni della L2 (Kuhl, Ramírez, Bosseler *et al.*, 2014). Hazan, Sennema, Iba & Faulkner (2005) hanno condotto uno studio per verificare se l'aggiunta di stimoli visivi nell'HVPT potesse portare a un miglioramento del riconoscimento fonetico rispetto all'uso di stimoli solo uditivi. Hazan *et al.* (2005) hanno usato stimoli audiovisivi che mostravano i movimenti delle labbra dei parlanti mentre producevano una serie di non-parole con i contrasti /v/-/b/ e /l/-/r/. I risultati dello studio hanno mostrato che gli stimoli audiovisivi portavano a un miglioramento significativo dell'identificazione delle consonanti nella coppia minima /b/-/v/, ma non nella coppia minima /l/-/r/. Questo risultato misto potrebbe essere dovuto alla natura più evidente dal punto di vista visivo dei suoni labiodentali e bilabiali della coppia /b/-/v/. Hazan, Sennema, Faulkner *et al.*, (2006) hanno indagato l'impatto della componente visiva sulla percezione del parlato da parte di parlanti L1 e L2 in ascolto degradato. Lo studio presentava frasi degradate con rumore, riverbero e compressione dell'ampiezza, con o senza l'input visivo corrispondente dei movimenti labiali di chi palava. I risultati hanno mostrato che sia i parlanti L1 madrelingua che L2 beneficiavano dei segnali visivi. Tuttavia, il beneficio variava a seconda del tipo specifico di degrado del suono e del *background* linguistico dei partecipanti (L1 o L2). I risultati suggeriscono che i segnali visivi possono svolgere un ruolo importante nell'aiutare la percezione del parlato in condizioni di ascolto difficili sia per i parlanti L1 che per quelli L2. Similmente, Hannah, Wang, Jongman, *et al.*, (2017) mostrano che in un compito di identificazione di sillabe in mandarino sia i parlanti L1 che L2 beneficiano delle informazioni visive.

Uno dei possibili motivi di questi risultati è che l'input visivo fornisce ulteriori segnali sui movimenti articolatori associati a specifici fonemi. Gli input visivi potrebbero quindi aiutare gli studenti L2 adulti a sviluppare una rappresentazione più accurata e sfumata del suono L2. Osservando un parlante madrelingua produrre un particolare suono, gli studenti L2 possono osservare i movimenti delle labbra, della lingua e della mascella necessari per produrre quel suono. Queste informazioni visive potrebbero aiutare gli studenti a sviluppare una rappresentazione più dettagliata e accurata del suono, facilitandone il successivo riconoscimento e produzione. Inoltre, alcuni fonemi L2 possono essere difficili da distinguere basandosi solo su segnali uditivi, ma possono essere distinti più facilmente se accompagnati da segnali visivamente salienti. Nel complesso, incorporare l'input visivo nell'apprendimento dei fonemi L2 potrebbe quindi fornire ulteriori segnali e informazioni che supportano la percezione e la produzione dei suoni della L2.

1.4 Lo studio presente

Lo studio di Hazan *et al.* (2005) ha dimostrato che l'uso di stimoli audiovisivi può migliorare l'identificazione dei suoni L2 in un paradigma HVPT. Tuttavia, non è ancora chiaro quale sia il ruolo delle specifiche informazioni articolatorio-visive nell'apprendimento fonetico della L2. Le informazioni articolatorio-visive, fornendo un supporto complementare e integrativo all'apprendimento fonologico uditivo, potrebbero infatti aiutare i parlanti L2 a percepire meglio le caratteristiche distintive dei fonemi della L2. L'apprendimento dei movimenti articolatori specifici dei fonemi della L2 può inoltre sviluppare una maggiore consapevolezza delle differenze fonologiche tra i fonemi L2 e quelli della L1. Inoltre, gli studenti devono consolidare le categorie fonologiche L2 apprese e generalizzarle in contesi diversi. In situazioni linguistiche immersive, questo avviene attraverso l'uso continuo della lingua e l'esposizione a una varietà di contesti linguistici. In questo senso, l'uso dell'HVPT, accoppiato all'uso delle informazioni articolatorie e uditive, potrebbe aiutare a generalizzare e consolidare le distinzioni fonologiche L2 oggetto di training. Per aumentare l'efficacia del training HVPT con stimoli audiovisivi, potrebbe essere utile utilizzare stimoli visivamente salienti come, ad esempio, l'apertura maggiore o minore della bocca o la presenza o meno dell'arrotondamento delle labbra. Per testare l'ipotesi che *features* visivamente salienti aiutino a consolidare le categorie fonologiche della L2, abbiamo utilizzato le informazioni articolatorie visive di due contrasti vocalici del Southern Standard British English (SSBE), /e-æ/ e /ʌ-ɒ/. Questi contrasti sono particolarmente difficili da discriminare uditivamente per i parlanti di italiano salentino (Escudero, Sisinni & Grimaldi, 2014), che è caratterizzato da cinque fonemi vocalici: /i, ɛ, a, ɔ, u/. La scelta di coppie minime specifiche, già testate uditivamente in altri studi sui parlanti L1 dell'italiano salentino, ci permette di verificare se l'aggiunta di informazioni articolatorie visive supporta lo sviluppo di rappresentazioni mnemoniche inerenti alle proprietà acustico-articolatorie dei contrasti fonologici L2. Secondo la teoria dei *Tratti Distintivi*, è ben noto che la coppia minima /e-æ/ basa la sua contrastività fonologica sul tratto acustico-articolatorio [±basso] (che, secondo il modello teorico, ha anche un ruolo nel processo percettivo): /e/ è caratterizzato dai tratti [-alto, -basso, -posteriore, -arrotondato], mentre /æ/ dai tratti [-alto, +basso, -posteriore, -arrotondato]. L'opposizione fonologica /ʌ-ɒ/ si fonda, invece, sui tratti [±basso] e [±arrotondato]: /ʌ/ è un fonema [-alto, -basso, +posteriore, -arrotondato], mentre /ɒ/ è [-alto, +basso, +posteriore, +arrotondato] (cfr. Odden, 2005).

In linea con il modello *Analysis by Synthesis* (Halle, 2001), il cervello di chi apprende un sistema linguistico è in grado di estrarre rappresentazioni audio-motorie dall'analisi del segnale acustico (integrando unità percettive e unità motorie); in questo modo, viene generata una mappa audio-motoria funzionale alla scoperta dei principi che oppongono i suoni linguistici fra loro in termini di tratti distintivi. Al contrario, secondo la *Motor Theory of Speech Perception* (Liberman, Mattingly, 1985) l'apprendimento di un sistema fonologico avviene

attraverso l'analisi e la memorizzazione dei gesti articolatori direttamente implicati nella generazione del segnale acustico. A generare le rappresentazioni mentali dei contrasti fonologici sarebbero, quindi, le configurazioni articolatorie astratte coinvolte nella produzione dei suoni percepiti: in questo modello, dunque, gli atomi dei suoni linguistici non sono i tratti distintivi, ma proprio le configurazioni articolatorie astratte.

L'apprendimento di contrasti fonologici della L2 in età adulta diventa un banco di prova per le due teorie. Con questo lavoro, infatti, ci proponiamo di verificare:

– Se l'utilizzo di training HVPT con informazioni articolatorie audiovisive chiaramente visibili migliora l'apprendimento percettivo dei suoni L2 da parte di parlanti L1 dell'italiano salentino rispetto al training solo uditivo.

In parallelo, il nostro obiettivo è anche testare gli assunti dei due modelli appena discussi all'interno dei processi apprendimento fonetico-fonologico di una L2 (limitatamente allo sviluppo percettivo), e quindi verificare:

– Se anche gli apprendenti adulti, come i bambini, fanno principalmente ricorso al segnale acustico per costruirsi rappresentazioni fonetico-fonologiche dei suoni L2, le informazioni visive non dovrebbero produrre nessun vantaggio percettivo. Al contrario, se, come ipotizza la *Motor Theory of Speech Perception*, sono i gesti articolatori a stimolare la generazione di rappresentazioni fonologiche, le informazioni audio-visive aggiunte dovrebbero facilitare la categorizzazione di quei fonemi L2 caratterizzati da specificità articolatorie come l'arrotondamento delle labbra.

2. *Metodo*

2.1 Stimoli

Abbiamo video-registrato 40 non-parole (20 per la coppia minima /e-æ/ e 20 per la coppia minima /ʌ-ɒ/) prodotte da sei madrelingua di SSBE (tre maschi e tre femmine) di età compresa tra 35 e 56 anni. Seguendo studi precedenti (e.g. Carlet, 2017), le vocali sono state prodotte in contesto CVC assieme a consonanti bilabiali occlusive e nasali (/p, b, m/), velari occlusive (/k, g/), labiodentali fricative (/f, v/), e dentali occlusive (/t, d/). Per elicitare la produzione delle non-parole contenenti i fonemi selezionato per lo studio, i partecipanti hanno pronunciato le non-parole all'interno di frasi del tipo "*it rhymes with bed* [pausa] *tet*". L'audio e il video sono stati registrati con l'ausilio di un microfono Speedlink collegato esternamente a una telecamera Sony HDR. La telecamera riprendeva la parte inferiore del viso del parlante, concentrandosi sui movimenti articolatori delle labbra e della lingua (vedasi Fig. 1, pannello Audiovisivo). Due parlanti madrelingua SSBE hanno giudicato su una scala da 1 a 5 le produzioni audiovideo, valutando la conformità delle non-parole con i fonemi del SSBE (scala da 1 non conforme - 5 totalmente conforme). Le non-parole che hanno ricevuto un grado di giudizio pari o superiore

a quattro sono state selezionate per il training. I videoclip e i file audio delle non-parole sono state estratte dalle frasi con il software VLC.

Per il gruppo di controllo sono stati usati degli esercizi di comprensione del testo seguiti da relative domande a scelta multipla. Gli esercizi sono stati realizzati per studenti di inglese L2 di livello CEFR B1 e sono stati tratti dal sito web https://test-english.com/reading/b1/.

2.2 Partecipanti e gruppi di training

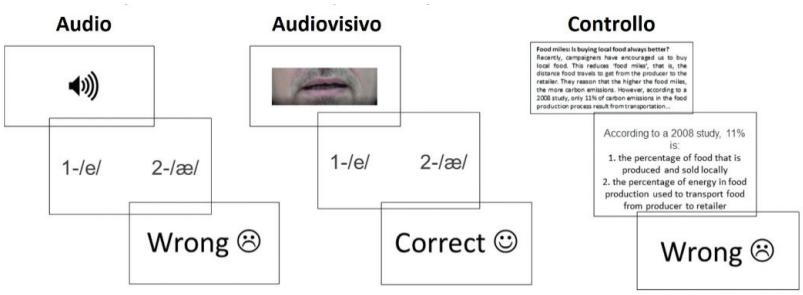
Allo studio hanno preso parte 66 parlanti di italiano salentino (età media dei partecipanti=19,5, M= 20, F= 46, livello CEFR inglese B1=51, A2=15). Tutti i partecipanti allo studio erano studenti del primo anno del corso di Lingue, culture e letterature straniere dell'Università del Salento e hanno ricevuto istruzioni riguardanti la fonetica acustica e articolatoria durante il corso di Linguistica generale che comprendevano anche le vocali prese in oggetto da questo studio. I parlanti selezionati erano tutti residenti in Salento da almeno 15 anni. Nessuno di loro era bilingue, e tutti avevano studiato inglese solo in contesto scolastico per almeno 8 anni, con soggiorni in paesi anglofoni non superiori ai 60 giorni. I partecipanti sono stati assegnati in modo casuale a tre gruppi sperimentali: training audio, audiovisivo o gruppo di controllo. I tre gruppi sono stati bilanciati per genere e livello di inglese.

2.3 Procedimento

Tutti i partecipanti hanno eseguito un pre-test uditivo di identificazione di 80 stimoli audio (40 non-parole x 2 parlanti nativi, 1 maschio e 1 femmina). Gli stimoli sono stati divisi in due sessioni di coppie minime (/e-æ/e /ʌ-ɒ/) in maniera randomizzata. Le sessioni sono state presentate ai partecipanti in modo controbilanciato (metà partecipanti hanno identificato prima i suoni della coppia minima /e-æ/e metà quelli della coppia minima /ʌ-ɒ/). Alla fine del pre-test i partecipanti hanno ricevuto un *feedback* globale sulla loro performance. I partecipanti assegnati ai gruppi di training audio e audiovisivo hanno preso parte a quattro sessioni di training ad intervalli settimanali, durante i quali hanno identificato 80 non-parole prodotte da 2 diversi parlanti (di sesso maschile e femminile) ricevendo *feedback* immediati (giusto/sbagliato) dopo ogni stimolo audio o audiovisivo identificato. Le non-parole sono state presentate ai partecipanti in due blocchi controbilanciati (un blocco per /e-æ/e un blocco per /ʌ-ɒ/). All'interno dei blocchi gli stimoli erano randomizzati. I partecipanti del gruppo di controllo hanno invece sostenuto a intervalli settimanali un test di comprensione del testo seguito da una serie di risposte a scelta multipla e feedback (Fig. 1). Gli stimoli erano presentati in modo randomizzato. Tutti i partecipanti hanno gestito il training in maniera autonoma, senza costrizioni temporali, ma generalmente la durata media di ogni sessione di training era di 25 minuti. Una settimana dopo la fine del training tutti i partecipanti hanno fatto un post-test uditivo di identificazione, con 40 non-parole, registrate da quattro

parlanti madrelingua, due (un maschio e una femmina) già ascoltati in precedenza durante il training e due nuovi parlanti (un maschio e una femmina) mai ascoltati in precedenza, e 16 nuove non-parole pronunciate sia dai due parlanti del training che dai due nuovi parlanti. Nel nostro studio, abbiamo addestrato i partecipanti utilizzando non-parole. Di conseguenza, per valutare la generalizzazione delle capacità percettive a nuovi suoni non inclusi nell'addestramento, abbiamo utilizzato nel post-test nuove non-parole.

Figura 1 - *Procedimento dei training audio, audiovisivo e del gruppo di controllo*



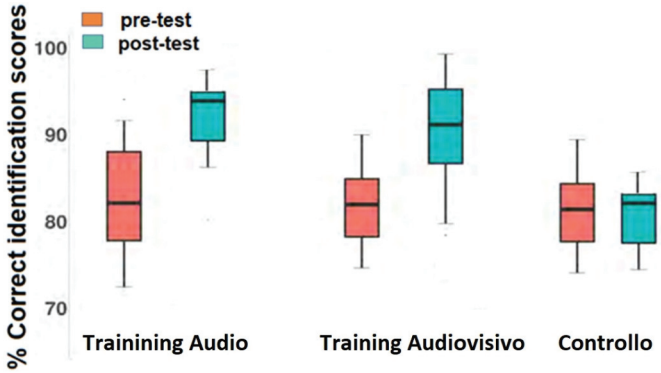
2.4 Analisi statistiche

I dati raccolti sono stati analizzati con una regressione logistica ad effetti misti, nella quale l'accuratezza dell'identificazione delle vocali era la variabile dipendente (1= vocale identificata, 0= vocale non identificata) e con il test (pre-test vs post-test), il metodo di training (audiovisivo, audio o gruppo di controllo) e le vocali /e/, /æ/, /ʌ/ e /ɒ/) come variabili indipendenti. Sono stati inoltre introdotti effetti casuali per partecipanti e item (non-parole). La regressione logistica è stata condotta con il software R versione 3.6 (R Core Team, 2022).

3. *Risultati*

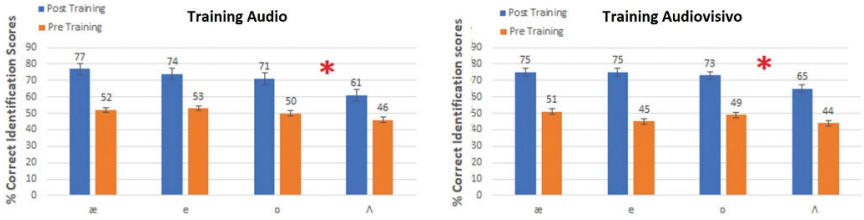
I risultati del modello statistico hanno mostrato una differenza significativa tra il pre-test e il post-test nell'identificazione delle vocali SSBE per i partecipanti al training audio rispetto ai partecipanti del gruppo di controllo (Est. = 0,77, SE=0,1, p<0,001) e tra il gruppo del training audiovisivo e il gruppo di controllo (Est. =0,97, SE=0,1, p<0,001) ma non tra i gruppi di training audiovisivo e audio (Est. = 0,2, SE=0,1, p=0,87, Fig. 2). In linea con risultati sul paradigma HVPT precedenti, non è stata trovata una differenza significativa né tra l'identificazione di nuove non-parole nel post-test e di non-parole usate durante il training (Est. =0.09, S. E=0.08, p=0.25) né tra l'identificazione di non-parole pronunciate dal nuovo parlante e dai parlanti ascoltati precedentemente durante il training (Est=0.04, S.E.=0.1, p=0.66).

Figura 2 - *Risultati del modello statistico: c'è una differenza significativa tra i risultati del pre-test e post-test nei partecipanti del gruppo audio e audiovisivo, ma non tra i partecipanti del gruppo di controllo. Inoltre, non ci sono differenze significative nei risultati del post-test tra i partecipanti del training audio e audiovisivo*



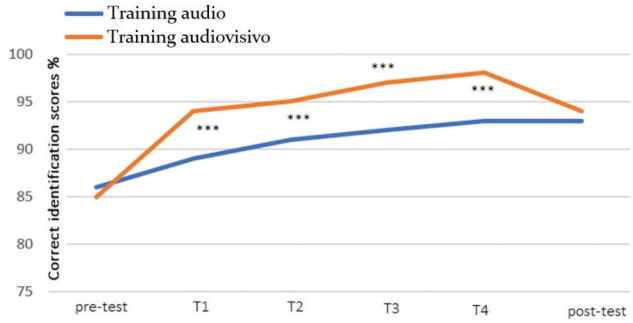
Esiste una differenza significativa tra l'accuratezza percettiva di /ʌ/ vs./ɒ/ sia per il gruppo del training audiovisivo (Est. = 0,96, SE=0,14, p<0,001) che per il gruppo che ha seguito il training audio (Est. =0,9, SE=0,14, p <0,001), suggerendo che per chi parla italiano salentino il contrasto vocalico SSBE più difficile da percepire è /ʌ/-/ɒ/ (Fig. 3). In particolare, la vocale /ʌ/ è quella che ha raggiunto i punteggi di identificazione più bassi tra tutte le vocali sia prima che dopo l'addestramento uditivo e audiovisivo.

Figura 3 - *Risultati del modello statistico: confrontando i risultati pre-training e post-training per i fonemi addestrati, si nota che il fonema /ʌ/ è quello significativamente più difficile da identificare (asterisco rosso)*



Un'ulteriore analisi statistica ha controllato se ci fosse una differenza significativa nell'identificazione dei fonemi del SSBE tra i partecipanti al training uditivo e audiovisivo. I risultati hanno dimostrato che i partecipanti del training audiovisivo identificano meglio i fonemi target rispetto ai partecipanti al traning solamente uditivo (Fig. 4, primo training (T1): Est=0.3, S.E.=0.1, p<0.001; secondo training (T2): Est=0.4, S.E.=0.1, p<0.001; terzo training (T3): Est=0.2, S.E.=0.1, p<0.001; quarto training (T4): Est=0.3, S.E.=0.1, p<0.001).

Figura 4 - *Confronto tra i risultati del pre-test (solo audio), training audio e audiovisivo a intervalli settimanali (T1= primo training, T2=secondo training, T3= terzo training, T4=quarto training) e post-test (solo audio). La percentuale di identificazione dei fonemi target è significativamente maggiore durante il training audiovisivo che in quello audio*



4. *Discussione e conclusioni*

I risultati dell'esperimento forniscono informazioni rilevanti rispetto alle ipotesi sperimentali elaborate. Con la prima ipotesi ci chiedevamo se l'utilizzo di training HVPT con l'aggiunta di informazioni articolatorie visive fosse più efficace nel migliorare l'identificazione dei suoni L2 del SSBE rispetto al solo training audio. In parallelo, rispetto a un processo di apprendimento dei suoni L2 in età adulta, ci siamo proposti di verificare alcuni assunti di base di due classici modelli teorici: l'*Analysis by Synthesis* e la *Motor Theory of Speech Perception* (cfr. 1.4).

I risultati hanno dimostrato che entrambi i gruppi sottoposti a training, sia quello audiovisivo che quello audio, hanno mostrato un miglioramento significativo nell'identificazione delle vocali del SSBE rispetto al gruppo di controllo. Tuttavia, non è emersa una differenza significativa tra i due gruppi di training audiovisivo e audio nel post-test. Questo suggerisce che l'aggiunta di informazioni articolatorie audiovisive non porta a un miglioramento nell'identificazione uditiva dei suoni L2 rispetto al solo training audio. Ciò sembrerebbe confermare gli assunti alla base del modello *Analysis by Synthesis*, secondo cui il parlante, anche adulto a questo punto, fa ricorso al segnale audio per generare rappresentazioni audio-motorie fonologicamente contrastive.

In particolare, i risultati del contrasto vocalico L2 /ʌ/-/ɒ/ dimostrano che – nonostante l'aggiunta delle informazioni articolatorie relative all'arrotondamento o meno delle labbra – i partecipanti hanno una percentuale di identificazione significativamente minore rispetto al contrasto /e-æ/, sia nel gruppo di training audiovisivo che nel gruppo di training audio. Inoltre, i bassi punteggi di identificazione della vocale /ʌ/ suggeriscono che i parlanti di italiano salentino hanno maggiori difficoltà nell'identificare questa vocale del SSBE, che viene confusa sistematicamente con la vocale /ɒ/.

Questo risultato richiede di essere adeguatamente interpretato, soprattutto in relazione ai modelli teorici precedentemente discussi. Nell'italiano salentino, che come abbiamo visto è caratterizzato da cinque fonemi vocalici, l'arrotondamento delle labbra – a differenza dei tratti [±alto] e [±posteriore] – non è un tratto distintivo che contribuisce alla contrastività fonematica. Infatti, il tratto [±arrotondato] in questo sistema è ridondante, in quanto le vocali /ɔ, u/, rispetto ad /i, ɛ/, sono entrambe [+posteriori] e [+arrotondate] e la vocale /a/ è caratterizzata solo dal tratto [+basso]: perciò il tratto [-arrotondato] di /a/ è predicibile. Dunque, il fatto che gli apprendenti adulti facciano ricorso al segnale acustico come risorsa privilegiata per generare rappresentazioni percettivamente contrastive permette loro di sviluppare una discriminazione fonemica solida per il tatto di altezza, come nel caso di /e-æ/, sfruttando un tratto già contrastivo nel sistema della L1; al contrario, lo stimolo acustico non sembra sufficiente per costruire una rappresentazione contrastiva robusta per il tratto [±arrotondato]. In altri termini, la risorsa articolatoria aggiunta nel training può agevolare il processo di identificazione, ma non sembra essere una informazione così saliente da riorientare la plasticità uditiva verso un contrasto fonologico L2 marcato da un tratto che nel sistema della L1 è ridondante.

Questa interpretazione ci pare supportata dai risultati riportati in Fig. 4. Da un lato, le informazioni audiovisive aggiunte producono un miglioramento del processo di discriminazione; dall'altro, le stesse informazioni (arrotondamento/non arrotondamento delle labbra) non sembrano sufficienti per generalizzare rappresentazioni fonemiche non presenti nel sistema contrastivo della L1. I nostri risultati ci sembrano in linea con quelli ottenuti da Hazan e colleghi (2005; 2006), i quali hanno evidenziato un miglioramento nelle capacità di discriminazione quando vengono fornite informazioni articolatorie audiovisive aggiuntive. Infatti, l'aggiunta delle informazioni articolatorie durante il training ha fornito ai partecipanti un vantaggio significativo nel riconoscimento dei suoni L2 rispetto al solo training audio. Potremmo dunque interpretare i nostri risultati come un supporto per l'ipotesi che l'attivazione dell'informazione motoria delle labbra giochi un ruolo nella percezione e nell'identificazione dei suoni del linguaggio, almeno durante la percezione audiovisiva. Questo potrebbe essere spiegato dal fatto che l'input visivo, che mostra i movimenti delle labbra dei parlanti, fornisce indizi motori aggiuntivi agli apprendenti. Il ricorso a questi indizi motori potrebbe quindi facilitare la corrispondenza tra l'input uditivo e la rappresentazione motoria del suono, migliorando così la percezione e l'identificazione dei suoni L2, ma non a tal punto, come nel nostro caso, da produrre la costruzione solida di rappresentazioni fonologiche distanti dalla L1. Infatti, quando nel post-test gli stimoli vengono presentati nella sola modalità audio emerge una netta discrepanza nella capacità di discriminare coppie di fonemi L2 differenziati rispetto ai tratti distintivi che hanno funzione contrastiva nel sistema della L1 ([±alto] vs. [±arrotondato]).

In sintesi, i risultati di questo studio indicano che sia il training audiovisivo che quello audio producono un miglioramento significativo nell'identificazione dei suoni L2 del SSBE per i parlanti L1 di italiano salentino. Il training audiovisivo,

dunque, non offre un vantaggio diretto nel rimodellare la plasticità uditiva degli apprendenti L2. E questo può dipendere dal fatto che, come abbiamo visto, i parlanti estraggono i parametri distintivi di un sistema fonologico direttamente dal segnale acustico (come assunto nel modello *Analysis by Synthesis*). Quando nel segnale acustico della L2 sono presenti informazioni relative a tratti distintivi contrastivamente già attivi nel sistema L1, queste vengono riutilizzate dagli apprendenti per costruire nuove rappresentazioni mnemoniche per i fonemi L2. Quando, invece, nel segnale acustico della L2 sono presenti informazioni relative a un tratto distintivo che nel sistema L1 non ha funzione contrastiva, gli apprendenti hanno difficoltà a costruirsi nuove rappresentazioni fonemiche stabili. Le ricerche future dovranno testare se in questi casi sia necessario un periodo di training più prolungato o se, in parallelo, associare al training uditivo un training articolatorio (aggiungendo, magari, un periodo di riflessione metalinguistica sulle proprietà fonetico-fonologiche dei suoni L2 che si stanno apprendendo). Infatti, se il bambino è in grado di scoprire in modo naturale le proprietà di qualsiasi sistema fonologico a cui è esposto in contesto naturale, l'adulto – come dimostrato da Kuhl e colleghi (2014) – mostra una differente attivazione del network neurale per avvicinarsi agli stessi obiettivi di apprendimento.

Nel complesso, questi risultati hanno importanti implicazioni per lo sviluppo di metodi didattici destinati agli adulti che apprendono una L2 in contesto formale. Nella progettazione e sperimentazione di training mirati (da utilizzare in futuro nella pratica didattica in classe) bisognerà, quindi, tenere nella giusta considerazione le differenze fonematiche fra il sistema L1 ed il sistema L2 e, in parallelo, pensare di sviluppare training flessibili in grado di adattarsi alle diverse proprietà della L2 rispetto alla L1 di partenza.

Bibliografia

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., Tohkura, Y.I. (1997). Training Japanese listeners to identify English /r/and /l/: IV. Some effects of perceptual learning on speech production. In *The Journal of the Acoustical Society of America*, 101(4), 2299-2310. https://doi.org/10.1121/1.418276

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. In *Language and Cognitive Processes*, 27(7-8), 1205-1224. http://dx.doi.org/10.1080/01690965.2012.660170

Carlet, A. (2017). L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study. PhD. Dissertation, Universitat Autonoma de Barcelona.

Cebrian, J., Carlet, A. (2014). Second-language learners' identification of target-language phonemes: A short-term phonetic training study. In *Canadian Modern Language Review*, 70(4), 474-499. https://doi.org/10.3138/cmlr.2318

Cutler, A., Weber, A., Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. In *Journal of Phonetics*, 34(2), 269-284. https://doi.org/10.1016/j.wocn.2005.06.002

Darcy, I., Dekydtspotter, L., Sprouse, R.A., Glover, J., Kaden, C., McGuire, M., Scott, J.H.G. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English– L2 French acquisition. In*Second Language Research*, 28, 5-40. http://dx.doi.org/10.1177/0267658311423455

Escudero, P., Hayes-Harb, R., Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. In *Journal of Phonetics*, 36, 345–360 https://doi.org/10.1016/j.wocn.2007.11.002

Escudero, P., Sisinni, B., Grimaldi, M. (2014). The effect of vowel inventory and acoustic properties in Salento Italian learners of Southern British English vowels. In *Journal of Acoustic Society of America*,135(3), 1577-84. https://doi.org/10.1121/1.4864477

Flege, J.E. (1987). A critical period for learning to pronounce foreign languages? In *Applied Linguistics*, 8(2), 162-177.

Flege, J. (1995a). Second language speech learning: Theory, findings, and problems. In Strange, W. (Ed.), *Speech perception and linguistic experience: Issues in cross language research*. Timonium, MD: York Press, 233-277.

Flege, J.E. (1995b). Two procedures for training a novel second language phonetic contrast. In *Applied Psycholinguistics*, 16, 425-442. https://doi.org/10.1017/S0142716400066029

Halle, M. (2001). *Form Memory to Speech and Back*. Berlin-New York: De Gruyter.

Hazan, V., Sennema, A., Iba, M., Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. In *Speech Communication*, 47(3), 360-378. https://doi.org/10.1016/j.specom.2005.04.007

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. In *Journal of Acoustoc Society of America*, 119(3), 1740–1751. https://doi.org/10.1121/1.2166611

Hannah, B., Wang, Y., Jongman, A., Sereno, J.A., Cao, J., Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. In *Frontiers in Psychology*, 8(2051). https://doi.org/10.3389/fpsyg.2017.02051

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. In *The Journal of the Acoustical Society of America*, 116(4), 2384-2394. https://doi.org/10.1121/1.1783351

Houde, J.F., Jordan, M.I. (1998). Sensorimotor adaptation in speech production. In *Science*, 279, 1213-1216. https://doi.org/10.1126/science.279.5354.1213

Iverson, P., Hazan, V., Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. In *The Journal of the Acoustical Society of America*, 118(5), 3267-3278. https://doi.org/10.1121/1.2062307

Iverson, P., Evans, G.B. (2009). Learning English vowels with different first-language vowel systems II: Auditor/ training for native Spanish and German speakers. In *Journal of Acoustical Society of America*, 126(8), 866-877. https://doi.org/10.1121/1.3148196

Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lin J.L., Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. Proceedings of the National Academy of Sciences. 111(31), 11238-11245. https://doi.org/10.1073/pnas.1410963111

Lengeris, A., Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. In *The Journal of the Acoustical Society of America*, 128(6), 3757-3768. https://doi.org/10.1121/1.3506351

Liberman, A.M., Mattingly, I.G. (1985). The motor theory of speech perception revised. In *Cognition*, 21(1), 1–36.

Lively, S.E., Logan, J.S., Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/, II: The role of phonetic environment and talker variability in learning new perceptual categories. In *The Journal of the Acoustical Society of America*, 94(3), 1242-1255.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. In *Journal of the Acoustic Societyof America*, 96(4), 2076-2087. https://doi: 10.1121/1.410149

Logan, J.S., Lively, S.E., Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/and/1/: A first report. In*The Journal of the Acoustical Society of America*, 89(2), 874-886.https://doi.org/10.1121/1.1894649

Major, R.C. (1987). Foreign accent: Recent research and theory. In*International Review of Applied Linguistics in Language Teaching*, 25(4), 185-202. https://doi.org/10.1515/iral.1987.25.1-4.185

Nishi, K., Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. In *Journal of Speech, Language, and Hearing Research*, 50(6), 1496-1509. https://doi.org/10.1044/1092-4388(2007/103)

Nishi, K., Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects ofvowels in sets and order of training. In *Journal of Speech, Language, and Hearing Research*, 51(6), 1480-1493. https://doi.org/10.1044/1092-4388(2008/07-0109)

Odden, D. (2005). *Introducing Phonology*. Cambridge: Cambridge University Press.

Pruitt,S., Jenkins, J.J., Strange, W. (2006). Training the perception ofHindi dental and retroflex stops by native speakers of American English and Japanese. In *The Journal of the Acoustical Society of America*, 119(3), 1684-1696. https://doi.org/10.1121/1.2161427

Qian, M., Chukharev-Hudilainen, E., Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. In *Language Learning and Technology*, 22(1), 69-96. https://dx.doi.org/10125/44582

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rato, A., Rauber, A. (2015). The effects of perceptual training on the production of English vowel contrasts by Portuguese learners. In the Scottish Consortium for ICPhS 2015. (Ed.), *Proceedings of the i8th International Congress of Phonetic Sciences*. Glasgow, UK: Glasgow University, 56.

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., Munhall, K.G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. In *The Journal of the Acoustical Society of America*, 123(1), 397-413.https://doi.org/10.1121/1.2804942

Thomson, R.I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. In *CALICO Journal*, 28,744-765.

Thomson, R.I. (2012). Improving Li listeners' perception of English vowels: A computer-mediated approach. In *Language Learning*, 62, 1231-1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R.I. (2016a). Does training to perceive Li English vowels in one phonetic context transfer to other phonetic contexts? In *Canadian Acoustic*, 44(3), 198-199.

Thomson, R.I. (2016b). Learning to perceive English sounds using computer-assisted pronunciation training. *AL Forum*, November 2016, Alexandria, VA, TESOL International Association.

Thomson, R.L., Derwing, T.M. (2016). Is phonemic training using nonsense or real words more effective? In Levis, J., Le, H., Lucic, I., Simpson, E., Vo, S. (Eds.). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*. Ames IA: Iowa State University, 88-97.

Wang, X., Munro, M.J. (2004). Computer-based training for learning English vowel contrasts. In*System*, 32(4), 539-552.

Wang,Y., Spence, M.M., Jongman, A., Sereno, J.A. (1999). Training American listeners to perceive Mandarin tones. In *The Journal of the Acoustical Society of America*, 106(6), 3649-3658. https://doi.org/10.1121/1.428217

Weber, A., Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. In *Journal of Memory and Language*, 50, 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

DOI: 10.17469/O2111AISV000022

NICHOLAS NESE

Arabo lingua straniera. Acquisizione di nuove categorie fonologiche

Arabic foreign language. Acquisition of new phonological categories

In this study, we investigate the acquisition of Arabic pharyngealized [tˤ], [dˤ], [sˤ], [ðˤ], pharyngeal [ħ], [ʕ], and uvular [q] sounds by Italian L1 learners. These sounds are widely acknowledged as challenging to learn (Mion, 2018; Shehata, 2018; Al Tubuly, 2018), both in terms of production and perception. 64 university students from 13 Italian universities, exhibiting varying proficiency levels, participated in an online experiment conducted through the Gorilla Experiment Builder platform. The experiment encompassed a word reading task alongside two perceptual tasks: identification and AX discrimination. The findings indicate improved performance with increased learner competence, revealing differing levels of difficulty among the sounds in both production and perception.

Keywords: Arabic, L2 acquisition.

1. *Introduzione*

Negli ultimi anni la lingua araba è stata oggetto di diversi studi volti a indagarne l'acquisizione fonetico-fonologica, sia a livello segmentale che soprasegmentale.

Eads, Khater, Mielke (2018) hanno condotto uno studio fonetico-articolatorio, indagando la sola produzione per mezzo dell'*ultrasound image*. I ricercatori hanno testato un gruppo di apprendenti di L1 inglese, con livelli di *proficiency* diversi, al fine di indagare le realizzazioni dei foni faringalizzati [tˤ] [dˤ] [sˤ] [ðˤ], uvulari [q] [χ] [ɣ], faringali [ħ] [ʕ] e glottidali [h] [ʔ] in diversi contesti fonologici. I dati raccolti hanno dimostrato che gli apprendenti principianti, ovvero quelli del primo anno, non sono in grado di distinguere l'articolazione dei foni non faringalizzati da quella dei foni faringalizzati o degli altri foni gutturali[1], con l'unica eccezione di quelli glottidali. Si ipotizza inoltre che, per gli apprendenti di L1 inglese, vengano acquisiti prima i foni glottidali, le consonanti faringalizzate sorde e l'occlusiva uvulare; successivamente le fricative uvulari e faringali, nonché le faringalizzate sonore.

Shehata (2018) ha esplorato la relazione tra produzione e percezione, testando un gruppo di apprendenti L1 inglese, con diversi livelli di *proficiency*, su dieci contrasti consonantici dell'arabo ma solo in posizione iniziale di parola: /t-tʕ/, /d-dʕ/, /θ-ð/, /ð-ðʕ/, /s-sʕ/, /h-ħ/, /k-q/, /ʔ-ʕ/, /χ-ɣ/ e /ħ-ʕ/. I dati raccolti hanno

[1] L'etichetta 'gutturale' viene comunemente utilizzata per identificare quel gruppo di suoni dell'arabo composto dalle uvulari, faringali, glottidali e faringalizzate.

rivelato una correlazione tra produzione e percezione. Inoltre, il numero di anni di studio della lingua araba è risultato essere significativo per gli apprendenti di livello avanzato, in quanto ha avuto un effetto positivo sia sulla produzione sia sulla percezione di tutti i contrasti.

Analogamente, Al Tubuly (2018) ha indagato la produzione e percezione dei foni consonantici faringalizzati, faringali e uvulari, ma con una prospettiva interlinguistica, reclutando cinque gruppi di apprendenti con L1 diverse: inglese, tedesco, greco, turco e cinese. Lo studio ha permesso di identificare i suoni faringalizzati e faringali come più difficili per gli apprendenti delle diverse lingue, sia sul piano della produzione che della percezione; laddove i suoni fricativi uvulari risultano essere quelli prodotti e percepiti con accuratezza migliore. Rispetto al confronto tra i diversi gruppi, invece, gli apprendenti turchi sono risultati i migliori; greci e cinesi si sono posizionati rispettivamente penultimi e ultimi. I risultati ottenuti fornirebbero dunque un supporto all'ipotesi secondo cui la L1 avrebbe un effetto sulla produzione e percezione della L2, almeno per alcuni suoni consonantici: infatti, le diverse produzioni non target degli apprendenti variano sulla base della presenza o assenza di certi tratti fonologici nelle rispettive L1.

Sempre sul piano segmentale, ma con attenzione rivolta a fenomeni del vocalismo, va menzionato lo studio di Zaltz, Segal (2021) che indagano la capacità di discriminare vocali lunghe e brevi in un gruppo di informanti con L1 ebraico.

Rispetto al piano soprasegmentale, nello studio di Al-Aloula (2018) è stata indagata la produzione di un gruppo di apprendenti di L1 inglese-americano, al fine di verificare l'effetto della L1 sulla produzione dell'accento lessicale in arabo L2. È emersa una differenza tra apprendenti e parlanti nativi, rispetto alla durata (ms), al pitch (Hz) e all'intensità (dB) nell'individuazione dell'accento: nella fattispecie, gli apprendenti di arabo L2 usavano la durata in misura maggiore rispetto ai parlanti nativi di arabo L1.

Sebbene dallo spoglio della letteratura siano emersi diversi studi volti a indagare l'acquisizione dell'arabo L2/LS, non ci risulta siano state condotte ricerche su apprendenti italofoni[2]. Questo lavoro è volto a colmare tale lacuna e si propone di indagare l'acquisizione di un gruppo di suoni consonantici dell'arabo, sia a livello di produzione sia di percezione, in apprendenti italiani L1 e con diverso grado di esperienza.

2. *La lingua araba*

Come è noto, l'arabo è una lingua semitica appartenente alla macro-famiglia linguistica camito-semitica (o afro-asiatica)[3]. Si tratta della lingua ufficiale dei 22 Stati membri della Lega araba nonché una delle sei lingue ufficiali delle Nazioni Unite. Occorre precisare che quando ci si riferisce all'arabo come lingua ufficiale

[2] Esaminando una prospettiva inversa, ovvero quando l'italiano è la L2 di apprendenti arabofoni, è rilevante menzionare lo studio condotto da Mori (2007) *Fonetica dell'italiano L2. Un'indagine sperimentale sulla variazione nell'interlingua dei marocchini.*

[3] Greenberg (1963: 42). Per una rassegna delle lingue camito-semitiche si veda Banfi, Grandi (2008).

di un Paese, nel contesto occidentale vengono adottate diverse definizioni (Mion, 2010: 190). In italiano è possibile imbattersi in 'arabo letterario moderno' (Manca, 1989), 'arabo classico' (Durand, 2009) o 'arabo standard moderno' (Mion, D'Anna, 2021)[4]. In tedesco, oltre a *klassisches Arabisch* (Fischer, 1972), viene adottata la dicitura *Hocharabisch* (Schultz, 2004). In francese viene contrapposto l'*arabe classique* (Larcher, 2012) all'*arabe littéral* (Neyreneuf, Al-Hakkak, 1996). In inglese viene fatta una distinzione fra *Classical Arabic* e *Modern Standard Arabic* (Elmahdy, Gruhn, Minker, 2012): quest'ultimo costituirebbe la diretta continuazione e versione moderna del primo (Fischer, 2011: 404). Si tratta infatti della stessa e identica lingua, sia sul piano strutturale che formale, in quanto condividono la stessa fonologia e morfosintassi (Mion, 2016: 45). Tale lingua nella tradizione araba viene identificata con una sola definizione: *al-ʿarabiyya l-fuṣḥā* "la [lingua] araba eloquentissima" o semplicemente *al-fuṣḥā* "la [lingua araba] eloquentissima" (Mion, 2010: 190). A livello sociolinguistico, la lingua araba presenta una situazione di diglossia: l'arabo standard (ovvero la *fuṣḥā*) coincide con la varietà alta, mentre il dialetto (*ʿāmmiyya*) con quella bassa[5]. Sul piano acquisizionale, il dialetto costituisce la L1 di tutti gli arabofoni (Kallas 2005: 151) laddove l'arabo standard rappresenta una varietà appresa in contesto scolastico. Dato che in Italia l'insegnamento della lingua araba in ambito universitario prevede principalmente lo studio della varietà standard[6], si è deciso di indagare dei fenomeni consonantici presenti nell'inventario fonologico di questa varietà.

2.1 Fonologia dell'arabo standard

L'inventario fonologico dell'arabo standard è composto da tre vocali (*ḥaraka*), a cui si aggiungono due dittonghi, e da ventotto consonanti (*ḥarf*). Le vocali si distinguono tra lunghe o brevi mentre le consonanti possono essere scempie o geminate. Si precisa che il sistema qui proposto (Tab. 1 e Tab. 2) è ampiamente condiviso e accettato dalle comunità arabofone come modello di pronuncia; tuttavia, alcuni suoni possono essere soggetti a variazione a seconda del dialetto specifico del parlante (Mion, D'Anna, 2021: 3).

Ai fini della presente ricerca è importante sottolineare che i principali manuali di arabo L2/LS adottati in Italia (e presumibilmente anche all'estero) e le grammatiche prevedono una descrizione dei suoni dell'arabo, più o meno esplicita[7], coerente con

[4] La differenza che intercorre fra queste definizioni riguarda l'evoluzione storica della lingua: con 'arabo classico' ci si riferisce alla scrittura letteraria adoperata da poco prima dell'avvento dell'Islam fino alla metà del XVIII secolo; con 'arabo moderno' si intende quell'evoluzione della *fuṣḥā* in senso moderno a partire dal XIX secolo (ovvero dalla Nahḍa); l''arabo standard moderno' rappresenta la fase contemporanea della *fuṣḥā* (Mion, D'Anna 2021).

[5] Nei lavori di arabistica specialistici si fa riferimento al dialetto utilizzando la dicitura di neoarabo (Mion 2016: 46).

[6] Lo studio dell'arabo standard viene solitamente integrato dallo studio di una varietà dialettale in una fase successiva. Modalità di insegnamento e varietà dialettale insegnata cambiano tra i diversi Atenei.

[7] In Mion & D'Anna (2021) per ogni suono viene indicato sempre il simbolo IPA e la descrizione puntuale del modo e luogo di articolazione. Talvolta viene fornita una descrizione articolatoria del

il sistema fonologico qui riportato. Per questi motivi, nel presente studio si preferisce adottare la dicitura di "suono target" invece di "suono standard", individuando i target di riferimento negli esiti riportati e descritti nella manualistica.

2.1.1 Vocalismo

Tabella 1 - *Vocalismo dell'arabo standard*

Vocali		*Dittonghi*	
		aj	aw
i i:	u u:		
a a:			

Le tre vocali presenti in arabo standard sono /a/ /i/ /u/ che, come già detto, possono essere brevi o lunghe; a queste si aggiungono i due dittonghi /aj/ /aw/. La quantità vocalica ha valore distintivo sia quando la vocale è tonica sia quando è atona (Mion & D'Anna 2021: 10). Esistono almeno due allofoni contestuali per ciascuna vocale: uno quando la vocale è in posizione marcata, ovvero se a contatto con una consonante uvulare, faringale, glottidale o faringalizzata; l'altro si presenta nel caso in cui la vocale sia in posizione non marcata, vale a dire se adiacente a un'altra consonante (Mion & D'Anna 2021: 11). Analogamente, i dittonghi in contesto marcato sono realizzati foneticamente come [aj] [aw], mentre quelli in posizione non marcata come [æj] [æw]. Nei dialetti, per le vocali lunghe non ci sono cambiamenti rilevanti, ad eccezione della quantità in sillaba tonica; mentre per le vocali brevi scompaiono in fine di parola (Durand 2009).

2.1.2 Consonantismo

Tabella 2 - *Consonantismo dell'arabo standard*

	Occlusive	*Fricative*	*Affricate*	*Laterali*	*Vibranti*	*Nasali*	*Approssimanti*	*Faringalizzate*
Labiali	b					m		
Labiodentali		f						
Dentali	t d			l	r	n		tˤ dˤ
Interdentali		θ ð						ðˤ
Alveolari		s z						sˤ
Postalveolari		ʃ	ʤ					
Palatali							j	

suono, ad esempio per [q]: "Per produrre questo suono la parte posteriore della lingua dovrebbe venire a contatto con l'ugola" (Ladikoff Guasto 2002: 51). In alternativa, i suoni arabi vengono comparati a quelli presenti in parole italiane o di altre lingue europee, ad esempio: "il suono *ch* dell'italiano *che*" (Tresso 1997: 14) oppure "come il *th* dell'inglese *think*" (Ladikoff Guasto 2002: 37).

	Occlusive	*Fricative*	*Affricate*	*Laterali*	*Vibranti*	*Nasali*	*Approssimanti*	*Faringalizzate*
Velari	k							
Labiovelari							w	
Uvulari	q	χ ʁ						
Faringali		ħ ʕ						
Glottidali	ʔ	h						

Sono presenti i suoni bilabiali /b/ /m/, l'approssimante labiovelare /w/e la labiodentale /f/, tutti comuni all'italiano. In arabo la sonorità è un tratto distintivo; ciò nondimeno, sono assenti l'occlusiva bilabiale sorda /p/, corrispettiva opposta della sonora /b/, e la fricativa labiodentale sonora /v/, corrispettiva opposta della sorda /f/. Mion (2010: 40) evidenzia che [p] può essere realizzato come allofono di [b] in contesti consonantici sordi; così come [v], in contesti sonori, può presentarsi come allofono di [f]. Durand (2009: 241-42) fa notare che, nei dialetti mesopotamici, in maltese e nei dialetti magrebini, il fono [p] ha acquisito valore fonologico; mentre [v] può presentarsi in alcuni europeismi.

Sono presenti i suoni dentali /t/ /d/ /r/ /l/ /n/ e /dˤ/ /tˤ/[8]. Come per le labiali, tutte le consonanti dentali dell'arabo si ritrovano anche in italiano ad eccezione dei suoni faringalizzati, assenti anche nelle altre lingue europee. Sono incluse nell'inventario due occlusive /t/ /d/ rispettivamente sorda e sonora, a cui corrispondono le due faringalizzate /tˤ/ /dˤ/. L'esito sordo /t/ può essere reso come affricata dentale sorda [ts] in alcuni dialetti parlati in Marocco centro settentrionale o in Algeria (Durand 2009: 239); l'autore precisa che tale esito si conserva solo in parlanti anziani. Vi è inoltre la laterale /l/ che viene realizzata come faringalizzata [lˤ] nella sola parola Allāh. (Mion 2016: 84). La rotica è polivibrante /r/, ma è possibile incontrare la realizzazione monovibrante [ɾ] come segnalato da Newman (2002). Tale esito, tuttavia, è da intendere come realizzazione allofonica (Mion 2010: 41). È da considerare dialettale l'esito uvulare [ʁ], che si può incontrare a Baghdad, Iraq settentrionale e nei dialetti del Tigri (Durand 2009: 238). È infine compresa nel raggruppamento di suoni dentali anche la nasale /n/.

Sono presenti i suoni interdentali /θ/ /ð/ e /ðˤ/. La fricativa sorda /θ/ e sonora /ð/ sono entrambe assenti in italiano ma presenti ad esempio in inglese; la realizzazione faringalizzata sonora /ðˤ/ è assente invece nelle lingue europee. Mion (2010: 41) fa notare che le interdentali presentano un'alta variabilità anche in arabo standard: si verifica un'occlusivizzazione nel Maghreb occidentale, con esiti [t] [d] [dˤ]; mentre nelle zone del Vicino Oriente si ha una sigmatizzazione con esiti [s] [z] [zˤ]. Durand (2009: 228) suggerisce che questo secondo fenomeno, in particolare in parlanti di varietà urbane, sia da attribuire alla forte stigmatizzazione dei suoni [θ] [ð] [ðˤ] percepiti come esiti rurali.

[8] De Dominicis (2013: 261) classifica /t/ /d/ /dˤ/ /tˤ/ come alveodentali.

Sono presenti le sibilanti /s/ /z/ e /sˤ/, di cui /s/ /z/, rispettivamente sorda e sonora, sono entrambe presenti in italiano mentre la realizzazione faringalizzata /sˤ/ è assente in tutte le lingue europee. In arabo non emerge variabilità areale.

Tra i suoni postalveolari e palatali sono presenti la fricativa sorda /ʃ/, l'affricata sonora /ʤ/ e l'approssimante /j/, tutti fenomeni che ritroviamo anche in italiano. Rispetto ai primi due segmenti consonantici, il sistema fonologico dell'arabo standard risulta essere asimmetrico: non è prevista né /ʒ/, realizzazione sonora corrispondente a /ʃ/, né /ʧ/ realizzazione sorda corrispondente a /ʤ/. Va precisato che /ʤ/ rappresenta un altro fonema soggetto a variazione, che viene realizzato come [ʒ] oppure [g] (quest'ultimo in Egitto) nell'oralizzazione dello standard (Mion 2010: 42). Un'elevata variabilità si registra anche sul piano dialettale. Duran (2009: 224) individua altre varianti oltre /ʤ/: [ʒ] rappresenta la realizzazione tipica dei dialetti urbani, dalla Siria al Marocco; [g] costituisce invece l'esito caratteristico del Cairo; [j] si ritrova nel sud dell'Arabia, nella costa orientale peninsulare e nella Mesopotamia meridionale; infine, /ʤ/ viene desonorizzata e resa come [ʧ] a Palmira.

Tra i suoni velari ed uvulari sono presenti /k/ /q/ /χ/ /ʁ/. L'occlusiva velare sorda /k/ è l'unico segmento comune all'italiano; l'occlusiva uvulare sorda /q/ è assente in italiano e in tutte le lingue europee; le fricative uvulari /χ/ /ʁ/, rispettivamente sorda e sonora, sono assenti in italiano ma presenti ad esempio in tedesco. L'opposizione sorda-sonora risulta parzialmente sbilanciata in quanto è presente solo per le fricative mentre per le occlusive, velari e uvulari, sono presenti solo gli esiti sordi. Durand (2009: 226) segnala esiti dialettali diversi da [k] nei dialetti beduini del Najd, della Palestina e della Giordania, dove viene realizzata come affricata dentale sorda [ʦ]; la realizzazione palatale [ʧ] è invece diffusa in numerosi dialetti beduini e rurali, peninsulari, palestinesi, iracheni e nordalgerini. Se a livello di arabo standard la /q/ non ammette varianti (Mion 2010: 42), sul piano dialettale sono quattro le realizzazioni che Durand (2009: 220) individua, oltre all'esito standard. Nelle zone siropalestinesi e nell'egiziano urbano viene realizzata come occlusiva glottidale [ʔ]. Viene realizzata come occlusiva velare sorda [k] nelle varietà rurali omanite, palestinesi e nordalgerine. L'esito occlusivo velare sonoro [g] risulta peculiare di tutti i dialetti beduini. Infine, viene realizzata come occlusiva uvulare sonora [ɢ] in Khūziztān. Per quanto riguarda le fricative /χ/ /ʁ/, va segnalato che spesso vengono classificate come velari [x][ɣ] (Thelwall & Sa'adeddin 1990).

Vi sono due fricative faringali /ħ/ /ʕ/ che costituiscono una coppia sorda-sonora. Entrambi i suoni sono assenti in italiano e in tutte le lingue europee[9]. Mion (2010: 44) fa notare che [ħ] presenta una maggiore costrizione rispetto a [ʕ]; di conseguenza è possibile imbattersi in realizzazioni di [ʕ] approssimanti, ad esempio in Siria e Vicino Oriente, e talvolta anche glottidali [ʔ], come in Iraq e Golfo (Mion & D'Anna 2021: 7).

[9] Costituisce un'eccezione il galiziano, in cui si segnala [ħ] come uno dei possibili esiti dialettali dell'occlusiva velare sonora (Regueira 1996: 120).

Sono presenti due glottidali, la fricativa /h/ e l'occlusiva /ʔ/, due fonemi assenti nell'inventario fonologico dell'italiano ma presenti entrambi in tedesco e, nel caso di /h/, anche in inglese. Tuttavia, in contesto italiano la [h] si ritrova nella gorgia toscana (Rohlfs 1966: 266), mentre il cosiddetto 'colpo di glottide' [ʔ] può essere realizzato in posizione iniziale davanti a vocale oppure in posizione intervocalica, in particolare nel parlato lento (Maturi 2006: 57). In arabo standard /ʔ/ risulta stabile laddove /h/ viene realizzato [ɦ] in contesti sonori (Canepari 2006: 104). A livello dialettale invece, Durand (2009: 230) segnala che /ʔ/ mantiene il suo valore fonologico in tutti i contesti (posizione iniziale, intermedia e finale) solo in alcuni dialetti peninsulari e in Yemen. Negli altri dialetti l'esito dipende soprattutto dal contesto fonologico: in posizione iniziale risulta facoltativo nei dialetti orientali, mentre scompare nei dialetti del Maghreb. In posizione intermedia viene sostituita da allungamento vocalico sia nei dialetti orientali sia occidentali; in posizione finale è sempre assente. Secondo Durand (2009: 231) l'elisione di [h] sarebbe un fenomeno ampiamente diffuso già nell'arabo magrebino medievale.

Infine, sono presenti nell'inventario fonologico dell'arabo standard quattro suoni faringalizzati: /tˤ/ /dˤ/ /sˤ/ /ðˤ/, noti come 'enfatiche'. Si tratta di suoni estranei sia all'italiano che alle lingue europee ma peculiari della lingua araba: l'arabo è infatti noto come 'lingua del ḍād', suono notoriamente complesso per i parlanti non arabofoni. Il termine 'enfasi' è stato adottato per la prima volta da Silvestre de Sacy nella sua *Grammaire arabe* (1810: 20)10 e da allora tale definizione è rimasta nell'uso. Una peculiarità delle lingue semitiche riguarda la presenza di consonanti apicali, tradizionalmente identificate come enfatiche, caratterizzate da una seconda articolazione, che può realizzarsi attraverso una velarizzazione, una faringalizzazione o una glottalizzazione (Mion 2010: 45). Vi è inoltre, nella sola parola Aḷḷāh, la laterale faringalizzata [lˤ], che è da ritenere quindi una realizzazione allofonica di [l] (Mion 2016: 84). Durand (2009: 236) segnala la presenza di [rˤ] nei dialetti orientali in presenza delle vocali /a/ /u/, in posizione pre o post consonantica, mentre Mion (2010: 47) ritiene che sia solo la vocale successiva ad avere un effetto sulla rotica. In arabo dialettale i suoni /dˤ/ e /ðˤ/ vengono quasi sempre confusi, confluendo quindi in una realizzazione unica. Nei dialetti in cui vengono mantenute le due interdentali /θ/ e /ð/ è presente solo /ðˤ/; laddove /θ/ e /ð/ non sono mantenute, se sostituite da /t/ e /d/ si ha /dˤ/, se realizzate come /s/ e /z/ si ha /zˤ/, se continuate da /f/ e /v/ si ha /vˤ/ (Durand 2009: 233). Durand fa inoltre notare che in dialetti periferici, ovvero in Asia centrale, Cipro, Malta o in Africa subsahariana, si perde la faringalizzazione.

[10] "Ce que j'appelle emphase ou articulation emphatique est une espèce de renflement qu'il n'est pas aisé de dèfinir mais qui fait en quelque sorte entendre un o sourd après la consonne: ainsi le mot داصن se prononce presque comme soad, sans cependant que cet o se fasse entendre distinctement".

3. *Modelli teorici acquisizionali*

A partire dagli anni Cinquanta[11] sono state avanzate diverse teorie acquisizionali, più o meno adeguate dal punto di vista descrittivo, esplicativo e predittivo a seconda del contesto di ricerca specifico, che si distinguono tra loro per aver posto l'attenzione su aspetti diversi[12] come la marcatezza (*Markedness Differential Hypothesis* di Eckman, 1977), la percezione fonologica dei tratti distintivi (*Phonological Interference Model* di Brown, 1998) oppure il piano articolatorio (*Articulatory Settings* di Honikman, 1964). Altri modelli sono stati sviluppati per essere applicati, almeno in origine, a specifiche situazioni socio-contestuali. Lo *Speech Learning Model* di Flege (1995), ad esempio, è stato pensato per apprendenti 'esperti' che vivono all'estero nel Paese della L2, laddove nel *Revised Speech Learning Model* (Flege, Bohn, 2021) cambia l'attenzione rispetto alla differenza tra *early learners* e *late learners*, non più oggetto di indagine. Il *Perceptual Assimilation Model* di Best (1995), inizialmente rivolto ad ascoltatori *naïve*, viene a sua volta ripensato per essere applicato ad apprendenti di L2 (Best, Tyler, 2007). A questi si aggiunge il *Second Language Perception Model* (L2LP) di Escudero (2005), modello specificamente progettato per tenere conto della variazione individuale in parlanti non madrelingua con un diverso livello di competenza.

4. *Domande di ricerca*

Come anticipato, la presente ricerca è volta a indagare l'acquisizione di un gruppo di suoni noti per essere difficili da acquisire (Mion 2018: 49; Shehata 2018: 56; Al Tubuly 2018: 70) sia a livello di produzione sia di percezione, ovvero i suoni faringalizzati [tˤ] [dˤ] [sˤ] [ðˤ] (noti anche come 'enfatiche'), faringali [ħ] [ʕ] e l'uvulare [q], in apprendenti con diverso grado di esperienza. Le principali domande di ricerca che hanno guidato il presente studio e a cui si cercherà di rispondere sono due:

1. Esiste una correlazione tra il livello degli apprendenti di arabo LS e le loro produzioni e percezioni di nuove categorie fonologiche?
2. Esistono dei fonemi e/o contrasti più difficili da acquisire nella produzione e/o nella percezione dell'arabo LS?

4.1 Ipotesi

A partire dalle domande di ricerca, sono state formulate delle ipotesi sui possibili risultati sulla base della MDH di Eckman, dello SLM-r di Flege e del PAM-L2 di Best.

Secondo lo SLM-r, tra i fattori che influenzerebbero la capacità di creare una nuova categoria figurano la quantità e qualità dell'input linguistico. Flege e Bohn, pur riconoscendo la difficoltà di quantificare l'input, propongono l'utilizzo della

[11] Lado (1957) è tra i primi a trattare in maniera sistematica l'acquisizione dei suoni della L2 all'interno della *Contrastive Analysis Hypothesis*, in un quadro improntato allo strutturalismo (Bloomfield 1933) e al comportamentismo (Skinner 1957) (Chini 2005: 20).

[12] Quella che segue è una rassegna solo parziale di modelli teorici. Non sono menzionati ad esempio fattori esterni che condizionano l'apprendimento come gli stili cognitivi, la personalità, la motivazione.

FTE (*full-time equivalent*), una misura che tiene conto della LOR (*length of residence*) e della proporzione dell'utilizzo della L2. Quindi, sulla base del modello proposto, ad un valore della FTE alto dovrebbero corrispondere prestazioni migliori. In riferimento al presente studio, si ritiene che il livello degli apprendenti possa costituire una misura sufficientemente buona della quantità e qualità dell'input ricevuto dagli apprendenti soprattutto in considerazione del fatto che il campione analizzato è estremamente eterogeneo. Pertanto, rispetto alla prima domanda di ricerca, l'ipotesi è che esista una correlazione positiva tra il livello degli apprendenti e le loro prestazioni, sia sul piano della produzione sia della percezione; ci aspettiamo quindi dei risultati migliori all'aumentare del livello degli apprendenti.

Sulla base della MDH di Eckmann, il grado di difficoltà sarebbe correlato alla marcatezza implicazionale del fenomeno da acquisire; di conseguenza il gruppo di suoni costituito da realizzazioni faringalizzate [sˤ] [dˤ] [tˤ] [ðˤ], faringali [ʕ][ħ] e dall'occlusiva uvulare [q], sarebbe difficile da acquisire. Tale previsione riguarderebbe il solo livello della produzione. La marcatezza intesa come implicazionale, tuttavia, non consente di formulare delle ipotesi sul grado di difficoltà dei suoni all'interno di questo gruppo, dal momento che non è possibile individuare delle relazioni tra di essi. Per poter avanzare delle ipotesi più precise su tutti i suoni sarebbe più adeguata, in questo caso, la definizione di marcatezza intesa come frequenza con cui un suono si presenta nelle lingue del mondo (Mairano *et al.* 2021: 13). In base a questo parametro e ai dati disponibili su UPSID e PHOIBLE, è possibile ipotizzare la seguente scala di difficoltà: [sˤ], [ðˤ], [dˤ], [tˤ] > [ʕ] > [ħ] > [q].

Nello SLM-r di Flege & Bohn, oltre alla quantità e qualità dell'input, si individua almeno un altro fattore da cui dipenderebbe la possibilità di creare una nuova categoria fonetica sia sul piano della produzione sia della percezione: la *Category Precision Hypothesis* (CPH), ovvero la precisione con cui il suono della L1 più vicino al suono target è definito all'interno dello spazio fonetico. Per quanto riguarda l'input, risulta difficile da utilizzare per formulare delle ipotesi rispetto al grado di difficoltà dei diversi suoni da acquisire. Tuttavia, non si può ignorare che l'esposizione ad una varietà dialettale possa influenzare sia la produzione che la percezione di un apprendente. Come già anticipato, la situazione sociolinguistica dell'arabo è caratterizzata da diglossia; il sistema fonologico dell'arabo standard, talvolta, non coincide con quello delle diverse varietà dialettali che però vengono utilizzate nella quotidianità. Ad esempio, [dˤ] è assente nell'inventario fonologico dell'arabo parlato in Iraq (Baghdad) e in Tunisia (Tunisi); [ð] e [ðˤ] risultano invece assenti nell'arabo parlato in Siria (Damasco), in Egitto (Cairo) e Marocco, come riportato da Mion (2010). È ragionevole quindi ipotizzare che realizzazioni dei partecipanti non aderenti allo standard e problemi nell'identificazione o discriminazione degli stimoli percettivi dell'arabo standard possano essere attribuiti, eventualmente, alla scarsa quantità e qualità di input ricevuto in fase di apprendimento. Rispetto alla CPH, Flege & Bohn definiscono la precisione di una categoria fonetica della L1 in termini di variabilità sul piano acustico delle produzioni relative a quella categoria fonetica. La metodologia adottata non consente di estrarre valori acustici utili ai fini

di una previsione; tuttavia, possiamo supporre che i partecipanti realizzino i suoni presenti nell'inventario fonologico dell'italiano con una variabilità acustica minore rispetto a quelli assenti. Quindi si potrebbe ipotizzare che [tˤ], [dˤ], [sˤ], [q] siano più facili da acquisire rispetto a [ðˤ], [ʕ], [ħ].

Seguendo il modello proposto da Best & Tyler (2007), è possibile formulare delle previsioni sul grado di difficoltà dei contrasti fonologici non nativi a livello percettivo. Tali previsioni tengono conto della distanza fonetica definita sulla base dei gesti articolatori; pertanto, nel contrasto [ʔ]-[ʕ] i suoni risultano essere più distanti rispetto agli altri, in quanto differiscono per due gesti articolatori, sia nel luogo che nel modo di articolazione. Questo contrasto rientra nella casistica numero 4, in cui si ipotizza una facile discriminazione dovuta all'associazione dei due suoni a set distinti di suoni della L1 e alla conseguente creazione di due nuove categorie fonologiche. Negli altri contrasti la differenza è determinata invece da un solo gesto articolatorio, luogo o modo di articolazione. Non è possibile determinare a priori se sia il luogo o il modo di articolazione a risultare più difficile da percepire; è però ragionevole pensare che la distanza fonetica nei contrasti [ðˤ]-[ð] e [h]-[ħ], entrambi costituiti da suoni non nativi, risulti più difficile da percepire rispetto a [t]-[tˤ], [d]-[dˤ], [s]-[sˤ], [k]-[q]. Quindi si ipotizza che i primi due contrasti rientrino nella casistica numero 3, secondo cui le categorie fonologiche della lingua target sono percepite come esempi equivalenti, ugualmente *good exemplar* o *poor exemplar*, della stessa categoria fonologica della L1; questo porta alla formazione di due categorie fonetiche equivalenti e difficili da discriminare, associate alla stessa categoria fonologica. Per quanto riguarda gli altri quattro, ipotizziamo che rientrino nella casistica numero 2, in cui due categorie fonologiche della L2 vengono percepite come equivalenti rispetto alla stessa categoria fonologica della L1, di cui una è percepita come esempio più distante rispetto alla categoria della L1; di conseguenza la difficoltà di discriminazione sarà moderata, almeno negli apprendenti di livello base. Entrambi i suoni saranno associati alla stessa categoria fonologica, ma a due categorie fonetiche distinte. Tuttavia, dal momento che uno dei due suoni sarà percepito più distante rispetto all'esito nativo, è possibile che si osservi una differenza tra studenti dei vari livelli e che quelli di livello avanzato riescano a discriminare meglio, in quanto potrebbero aver già sviluppato una nuova categoria fonologica.

5. *Metodologia*

Tradizionalmente la ricerca in ambito fonetico viene associata a contesti che prevedono la raccolta dati in presenza, sia che si tratti di studi di stampo sociofonetico, caratterizzati da indagini condotte sul campo, sia che si tratti di studi sperimentali condotti in laboratorio. Solo negli ultimi anni, anche a causa della pandemia di Covid-19, si è iniziato a esplorare nuove modalità di ricerca, ovvero a distanza, in modo da superare certi ostacoli legati all'impossibilità di condurre studi in presenza. Ed è proprio in questa seconda tipologia che si inserisce la presente ricerca. La raccolta dati di questo studio è stata condotta tra dicembre 2020 e

maggio 2021, in piena pandemia di Covid-19. Per tale motivo si è reso necessario individuare un tool che consentisse di svolgere l'esperimento a distanza.

Dopo aver vagliato le possibili soluzioni disponibili, si è deciso di utilizzare Gorilla (www.gorilla.sc), un tool sviluppato per la progettazione di esperimenti online e accessibile da settembre 2016. Grazie a un'interfaccia facile da usare, che consente anche a chi non ha competenze di programmazione di sviluppare esperimenti a distanza, e alla possibilità di realizzare sia task di produzione sia di percezione, Gorilla è risultato lo strumento ideale ai fini del presente studio.

5.1 Il campione

Hanno preso parte allo studio 125 studenti universitari di 13 Atenei del nord, centro e sud Italia. Ai fini di questa ricerca sono stati considerati complessivamente 64 partecipanti: sono stati esclusi dal campione coloro che hanno segnalato una L1 diversa dall'italiano (5 soggetti), nonché coloro che hanno indicato di avere disturbi uditivi (5 soggetti) o di eloquio (2 soggetti). Infine, si è deciso di non includere nel campione tutti coloro i cui audio registrati nel task di produzione risultavano disturbati o laddove si è riscontrato un elevato numero di stimoli mancanti. Questa scelta ha consentito un confronto ottimale dei risultati dei tre diversi task, dal momento che l'analisi ha interessato sempre lo stesso campione.

Tutti i partecipanti sono studenti universitari di età compresa tra i 19 e 27 anni (media 21.36, mediana 21.00, DS 1.785). È stato inoltre reclutato un gruppo di controllo costituito da 15 studenti universitari di età compresa tra i 18 e 24 anni (media 20.73, mediana 20.00, DS 2.017). Rispetto alla variabile di genere, il campione è costituito da 54 femmine (84 %), 9 maschi (14 %) e 1 agender (2 %); mentre il gruppo di controllo è composto da 11 femmine (73 %) e 4 maschi (27 %). È evidente lo sbilanciamento del campione che tuttavia risulta rappresentativo della realtà universitaria: la frequenza ai corsi di lingua straniera è spesso prevalentemente femminile. Tale distribuzione non ha consentito in fase di analisi un confronto dei risultati rispetto a questa variabile, motivo per cui è stato possibile includere l'unico soggetto che nel questionario si è dichiarato agender.

Considerato il campione di riferimento della ricerca, ovvero studenti universitari che studiano la lingua araba come LS, si è deciso di effettuare una classificazione dei livelli degli apprendenti sulla base delle informazioni curricolari (numero di corsi, CFU, semestri/annualità). La partecipazione allo studio da parte di studenti provenienti da diversi Atenei italiani ha richiesto una riflessione più attenta sui criteri da utilizzare. In primo luogo occorre considerare il fatto che l'insegnamento della lingua araba si inserisce in Corsi di laurea diversi, ad esempio: Lingue e Culture moderne (L-11), Mediazione Linguistica (L-12), Scienze della Comunicazione (L20), Scienze Politiche e delle Relazioni Internazionali (L-36), per quanto riguarda le lauree triennali; Lingue e Letterature dell'Africa e dell'Asia (LM-36), Lingue e Letterature Comparate (LM-37), Lingue Moderne per la Comunicazione e la Cooperazione Internazionale (LM-38), per le lauree magistrali. Da questa prima macro-distinzione deriva una pianificazione didattica differenziata, che in

alcuni casi prevede corsi semestrali laddove in altri i corsi sono annuali, anche a parità di ore o CFU: si possono infatti trovare corsi da 12 CFU distribuiti su un singolo semestre o sull'intera annualità. A ciò bisogna infine aggiungere la didattica integrativa con i docenti madrelingua, che può incidere significativamente sul totale delle ore di studio. Dopo aver valutato diverse opzioni, si è deciso di suddividere i partecipanti in tre gruppi: 16 studenti di livello base (25 %), 26 di livello intermedio (41 %), 22 di livello avanzato (34 %). L'assegnazione di ciascun partecipante ad un determinato gruppo è stata effettuata principalmente sulla base del numero di CFU conseguiti: da 8 a 12 CFU gruppo base; da 18 a 24 CFU gruppo intermedio; da 25 a 66 gruppo avanzato. Questa scelta ha consentito di ottenere dei raggruppamenti relativamente omogenei anche rispetto al numero di corsi frequentati.

5.2 L'esperimento

L'esperimento è stato composto da cinque moduli: un questionario iniziale, finalizzato all'inquadramento del partecipante; un task di lettura di parole in isolamento, per l'elicitazione di dati di produzione; un task di identificazione e uno di discriminazione, per l'elicitazione di dati percettivi; un questionario finale per ottenere un riscontro sullo svolgimento delle diverse prove. Si è deciso di collocare il task di lettura sempre prima dei due task percettivi per evitare che gli stimoli di questi ultimi potessero influenzare le produzioni dei partecipanti. Al contrario, i due task percettivi venivano presentati in ordine randomizzato: il task di identificazione poteva precedere quello di discriminazione AX e viceversa. In questo caso, la decisione è stata dettata dal voler evitare che fosse sempre lo stesso task ad essere svolto per ultimo. Si è potuto di conseguenza ridurre il rischio che l'eventuale stanchezza dovuta alla durata dell'esperimento potesse compromettere l'esito della prova e, qualora si fosse verificata tale situazione, si sarebbe potuta osservare attraverso un confronto dei risultati rispetto all'ordine di svolgimento delle prove.

5.3 Stimoli

Sono state stilate tre liste di parole distinte, una per ogni task. Quella utilizzata per il task di discriminazione AX era composta esclusivamente da coppie minime. Per ogni lista, sono state individuate 42 parole reali monosillabiche con il suono target in posizione iniziale seguito da vocale /a i u/[13]. Nei diversi task gli stimoli venivano presentati sempre randomizzati.

Per quanto riguarda il task di lettura si è deciso di presentare gli stimoli in caratteri arabi. La conoscenza della scrittura araba costituisce uno dei primi obiettivi previsti dai corsi di lingua base, pertanto tutti i partecipanti allo studio, anche quelli di livello base, sono stati in grado di svolgere la prova. Un possibile problema che è stato considerato nella fase preliminare riguardava la lettura vocalica della parola, ovvero come il partecipante avrebbe vocalizzato la parola target: in arabo, come anticipato,

[13] La vocale era quasi sempre lunga, tranne nei casi in cui non è stato possibile trovare delle parole reali con questa caratteristica.

la lunghezza vocalica ha valore distintivo e le vocali lunghe vengono sempre rese tramite grafemi mentre la presenza di vocali brevi può essere segnalata tramite diacritici. I testi in arabo però non sono quasi mai vocalizzati, pertanto, per leggere correttamente una parola con vocale breve occorre conoscerla. Al fine di renderne la lettura facilitata si è deciso quindi di vocalizzare tutte le parole con vocale breve.

Per quanto riguarda invece i task percettivi, gli stimoli sono stati registrati da cinque parlanti madrelingua di quattro nazionalità diverse (Marocco, Egitto, Tunisia, Giordania) di cui tre donne e due uomini, tutti con incarichi di docenza a livello universitario. Gli audio raccolti sono stati successivamente segmentati e normalizzati a 1 dB. La normalizzazione ha consentito l'utilizzo di tracce audio ottenute da registrazioni diverse che altrimenti non potevano essere ascoltate consecutivamente durante lo stesso task percettivo senza creare disturbo all'ascoltatore. Gli audio per il task di discriminazione AX sono stati quindi assemblati in modo che ogni stimolo contenesse le registrazioni di due voci diverse ma sempre dello stesso sesso. Per quanto riguarda l'intervallo tra gli stimoli, in linea con Escudero, Benders, Lipski (2009), è stato fissato a 1,2 secondi: Werker, Logan (1985) mostrano come un diverso intervallo fra gli stimoli può influenzare la modalità con cui l'ascoltatore classifica i suoni; più precisamente, un intervallo di 500 ms favorirebbe una comparazione acustica mentre un intervallo di 1.5 s garantirebbe un confronto fonologico.

6. *Analisi*

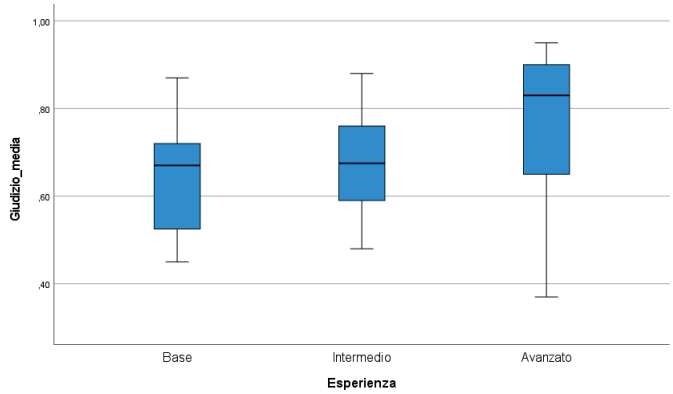
6.1 Task di produzione

L'analisi[14] del task di produzione è stata effettuata utilizzando come parametro la percentuale di risposte *target like* (TL). Ogni parola registrata dai partecipanti è stata valutata da chi scrive come accettabile (*target like*) o non accettabile (*non target like*), su base percettiva, rispetto alla realizzazione fonetica dell'arabo standard. È stata eseguita un'ulteriore valutazione di conferma sul 10 % del corpus da parte di due docenti di lingua araba, di cui uno arabofono madrelingua, con un agreement dell'85 %. Complessivamente le risposte TL sono risultate essere il 70 %. È stata successivamente integrata la variabile esperienza per verificare la variazione rispetto al livello degli apprendenti; i partecipanti di livello base hanno ottenuto una media del 64 % di risposte TL, il gruppo intermedio ha ottenuto una media del 67 % e quello avanzato del 77 %. È stata quindi condotta un'ANOVA: $F(2, 61) = 4.987$, $p < .010$, $\eta^2 = .141$. Il test ha rivelato una differenza statisticamente significativa tra i diversi gruppi e il valore dell'η^2 parziale ha evidenziato un *large effect* size[15]. È stato quindi eseguito un test post hoc di Tukey che ha rivelato una differenza statisticamente significativa tra il gruppo base e il gruppo avanzato ($p < .016$) e tra il gruppo intermedio e il gruppo avanzato ($p < .034$). Non è emersa significatività tra il gruppo base e intermedio ($p < .812$).

[14] Tutte le analisi qui riportate sono state eseguite tramite il software di analisi statistica SPSS 27.

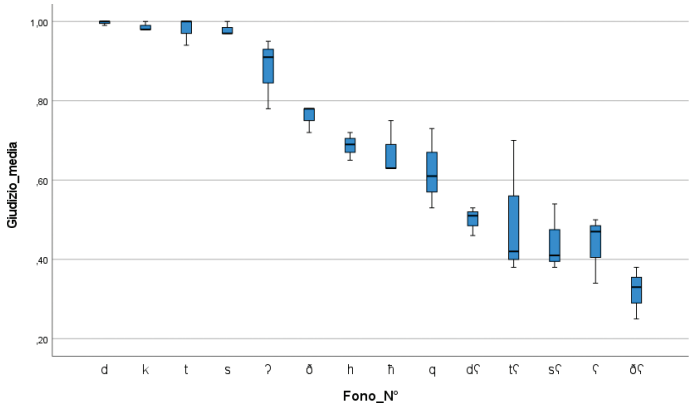
[15] Valori intorno allo .01 mostrano un *week effect*, quelli intorno a .06 un *medium effect* mentre un valore di circa .14 un *large effect* (Cohen 1988).

Figura 1 - *Percentuali degli esiti TL del task di produzione dei partecipanti suddivisi per esperienza*



Per ciascun fono è stata calcolata la media degli esiti valutati come realizzazioni TL. Si è provveduto ad eseguire un'ANOVA di Welch: $F(13, 27) = 29.404$ $p < .001$ che ha confermato la significatività statistica delle differenze fra i vari esiti. È sato quindi eseguito un test post hoc di Games-Howell, attraverso il quale è stato possibile individuare tre gruppi di foni etichettabili come difficili [ðˤ] [ʕ] [sˤ] [tˤ] [dˤ], medio-difficili [q][ħ][h] [ð], facili [ʔ][s][t] [k] [d].

Figura 2 - *Percentuale di esiti TL del task di produzione per ciascun fono*

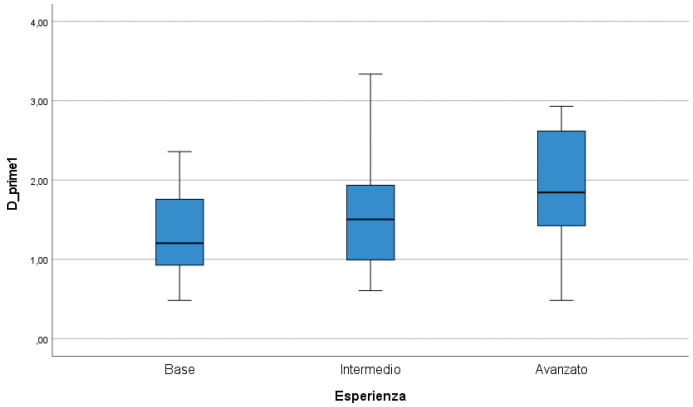


6.2 Task di identificazione

L'analisi del task percettivo di identificazione ha registrato complessivamente l'82 % di risposte TL, ovvero di identificazioni TL. È stato quindi calcolato per ogni partecipante il D prime. Per ciascuno dei tre gruppi, il valore medio è risultato essere di 1.3 per il gruppo di livello base, di 1.4 per il gruppo di livello intermedio e di 1.8 per il gruppo di livello avanzato. È stato condotto un test di Kruskal-Wallis per verificare se fosse possibile identificare una differenza tra i tre gruppi. Il test ($H(2) = 8.386$,

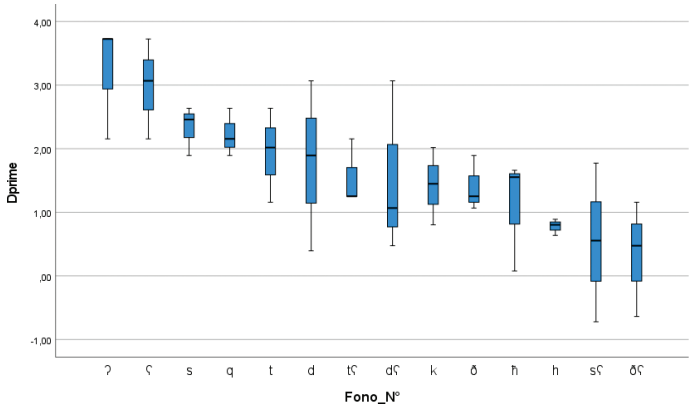
p< .015, a due code) ha rivelato una differenza statisticamente significativa tra i diversi livelli degli apprendenti, nello specifico tra il gruppo base e il gruppo avanzato (p< .010) nonché tra il gruppo intermedio e quello avanzato (p< .018). La differenza tra gruppo base e intermedio non è risultata statisticamente significativa (p< .608).

Figura 3 - *D prime del task di identificazione dei partecipanti suddivisi per esperienza*



Per ciascun fono è stata eseguita un'ANOVA utilizzando il D prime come variabile dipendente: F (13, 28) = 3.011, p < .007, η^2= .583. Il test post hoc di Tukey ha rivelato una differenza statisticamente significativa tra [ðˤ]- [ʕ] (p < .032), [ðˤ]- [ʔ] (p < .015) e [sˤ]- [ʔ] (p < .03).

Figura 4 - *D prime degli stimoli del task di identificazione per ciascun fono*



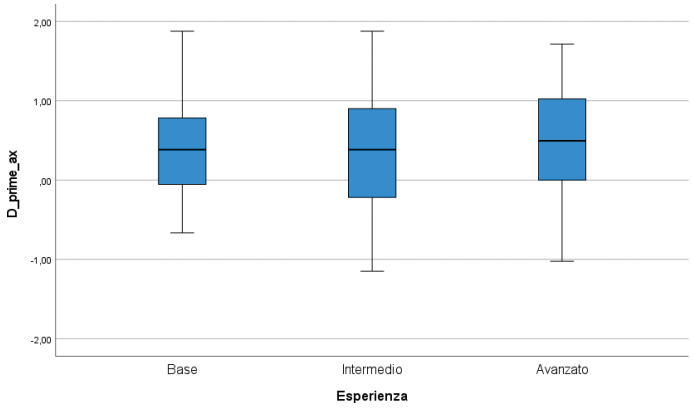
6.3 Task di discriminazione AX

Nel task di discriminazione AX è stata registrata una media complessiva di risposte TL del 65 %.

È stato quindi calcolato per ciascun partecipante il D prime; i valori medi di ciascun gruppo sono risultati essere 0.41 per gli apprendenti base, 0.37 per gli apprendenti di

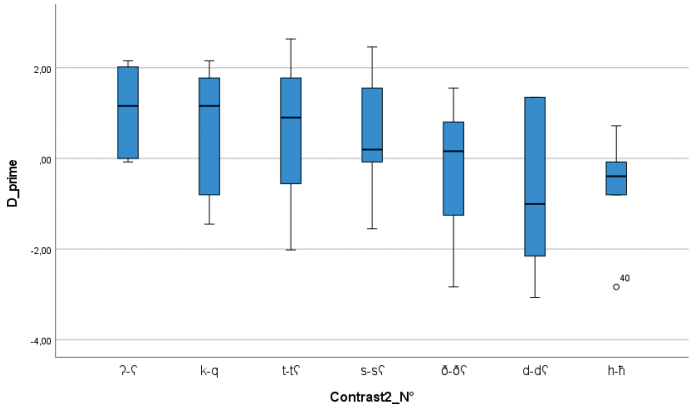
livello intermedio e 0.42 per quelli di livello avanzato. Sulla base di questo parametro è stata eseguita un'ANOVA F (2, 61) = .023, p < .978 che ha rivelato l'assenza di una differenza statisticamente significativa tra i tre gruppi di partecipanti.

Figura 5 - *D prime del task di discriminazione dei partecipanti suddivisi per esperienza*



Si è provveduto quindi al calcolo dei valori medi del D prime per ciascun contrasto e anche in questo caso l'analisi statistica non ha rivelato una differenza statisticamente significativa tra i diversi contrasti: ANOVA F(6, 35) =1.376, p < .252. Un'osservazione puntuale dei diversi stimoli utilizzati nel task ha rivelato[16] che negli stimoli contenenti parole diverse, laddove la prima parola iniziava con suono nativo (o comunque noto) gli stimoli erano sempre più facili da discriminare rispetto a quelli in cui la prima parola iniziava con un suono non nativo; costituisce l'unica eccezione il contrasto [ʔ]-[ʕ], in cui l'ordine delle parole non risulta avere un effetto sulla discriminazione.

Figura 6 - *D prime del task di discriminazione AX per contrasti*



[16] Tale osservazione non è stata sottoposta a test statistici.

7. *Discussione dei risultati*

Il contesto straordinario in cui si è svolta la ricerca, legato all'emergenza Covid-19, ha avuto un forte impatto sullo studio e in particolare sulle scelte metodologiche. Ciò non di meno, il corpus di dati raccolto e analizzato ha consentito di definire un quadro che, seppur preliminare, mette in luce alcuni fenomeni che aiutano a comprendere meglio – ma non del tutto – i processi acquisizionali di nuove categorie fonologiche dell'arabo LS da parte di apprendenti italofoni. Per questo motivo i risultati ottenuti e qui discussi sono da considerarsi non come un punto di arrivo ma piuttosto come un punto di partenza per possibili sviluppi futuri.

Rispetto alla prima domanda di ricerca, volta a verificare l'esistenza di una correlazione tra il livello degli apprendenti di arabo LS e le loro produzioni e percezioni di nuove categorie fonologiche, i risultati hanno rivelato che le performance del gruppo intermedio sono quasi sempre migliori rispetto a quelle del gruppo base e gli esiti del gruppo avanzato sono sempre migliori rispetto al gruppo intermedio e base. Tuttavia, la differenza tra i tre gruppi, in particolare tra il gruppo avanzato e i gruppi base e intermedio, è risultata statisticamente significativa solo per il task di lettura e di identificazione.

Sulla base dei risultati emersi, è possibile riconoscere una correlazione tra il livello degli apprendenti e le loro produzioni e identificazioni; questo risultato conferma quanto emerso dall'indagine di Shehata (2018), dove si evidenzia una differenza statisticamente significativa solo tra gli studenti di livello base-intermedio e gli studenti di livello avanzato. Per poter osservare un miglioramento significativo risulta quindi necessario considerare una lunga esposizione alla LS\L2, in linea con Flege, Bohn (2021) i quali sostengono che l'apprendimento di una L2 è un processo lento. Non risulta invece correlato il livello degli apprendenti con la capacità di discriminare i suoni.

Per quanto riguarda la seconda domanda di ricerca, volta a verificare se esistono dei fonemi e/o contrasti più difficili da acquisire nella produzione e/o nella percezione dell'arabo LS, sul piano della produzione è stato possibile definire il grado di difficoltà dei diversi suoni target dal più facile al più difficile: [ħ]> [q]> [dˤ]> [tˤ]> [sˤ] [ʕ]> [ðˤ]. Inoltre, l'analisi statistica ha permesso di individuare all'interno di questo gruppo degli esiti classificabili come medio-difficili ([ħ] [q]) e difficili ([dˤ] [tˤ] [sˤ] [ʕ] [ðˤ]). Né l'ipotesi formulata sulla base del modello di Eckman ([q]> [ħ]> [ʕ]> [sˤ] [tˤ] [dˤ] [ðˤ]) né quella relativa al modello di Flege, Bohn ([sˤ] [tˤ] [dˤ] [q]> [ħ] [ʕ] [ðˤ]) si è rivelata pienamente accurata, sebbene in entrambi i casi sia stata prevista una minore difficoltà del suono occlusivo uvulare sordo [q] rispetto a una difficoltà maggiore della fricativa interdentale sonora faringalizzata [ðˤ]. Non è quindi possibile stabilire se sia la marcatezza o la precisione delle categorie fonetiche della L1 ad avere un maggior valore predittivo.

Anche i risultati del task di identificazione hanno permesso di individuare una scala di difficoltà dei singoli segmenti, dal più facile al più difficile da identificare: [ʕ]> [q]> [tˤ]> [dˤ]> [ħ]> [sˤ]> [ðˤ]. Sulla base dei valori del D prime è stato possibile classificare i suoni in facili ([ʕ] [q]; D prime > 2), medio-difficili ([tˤ]

[dˤ] [ħ]; 2 > D prime > 1) e difficili ([sˤ] [ðˤ]; D prime < 1). L'analisi statistica ha inoltre evidenziato all'interno di questo gruppo una differenza significativa tra [ʕ] e [ðˤ], rispettivamente il suono più facile e più difficile da identificare. L'ipotesi formulata sulla base del modello di Flege e Bohn ([sˤ] [tˤ] [dˤ] [q]> [ħ] [ʕ] [ðˤ]) si è rivelata accurata solo parzialmente; è stata prevista solo una minore difficoltà del suono occlusivo uvulare sordo [q] rispetto a una difficoltà maggiore della fricativa faringale sorda [ħ] e della fricativa interdentale sonora faringalizzata [ðˤ].

Infine, i risultati del task di discriminazione AX hanno rivelato la seguente scala di difficoltà dei contrasti, dal più facile al più difficile da discriminare: [ʔ]-[ʕ]> [k]-[q]> [t]-[tˤ]> [s]-[sˤ]> [ð]-[ðˤ]> [h]-[ħ]> [d]-[dˤ]. Sulla base dei valori del D prime è stato possibile classificare i suoni in medio-difficili ([ʔ]-[ʕ]; D prime > 1), difficili ([k]-[q] [t]-[tˤ] [s]-[sˤ]; 1> D prime > 0) e molto difficili ([ð]-[ðˤ] [h]-[ħ] [d]-[dˤ]; D prime < 0). L'ipotesi formulata sulla base del modello di Best e Tyler ([ʔ]-[ʕ]> [k]-[q] [t]-[tˤ] [s]-[sˤ] [d]-[dˤ]> [ð]-[ðˤ] [h]-[ħ]) si è rivelata molto accurata. Il contrasto [ʔ]-[ʕ], caratterizzato da due suoni distanti tra loro e distanti rispetto alle categorie fonologiche dell'italiano, è risultato essere il meno complesso da discriminare per tutti gli apprendenti. Per il gruppo di controllo, invece, risulta il più difficile da discriminare. Questa situazione, apparentemente contraddittoria, confermerebbe quanto evidenziato nel PAM-L2 ovvero che il livello fonetico e livello fonologico possono influenzare la capacità di discriminazione in misura diversa e che pertanto vadano considerate separatamente. L'attribuzione di un valore fonologico a dei tratti fonetici, seppur molto distanti da quelli nativi, necessita una consapevolezza che dei non apprendenti non hanno.

8. *Prospettive di ricerca*

I risultati ottenuti nel corso di questa ricerca rappresentano un ottimo punto di partenza per possibili indagini future. L'ampia portata dello studio, che si basa su un campione ampio ed eterogeneo, ha permesso di elaborare delle considerazioni generali che risultano adeguate per identificare dei nuovi filoni di ricerca da approfondire.

Si potrebbe innanzitutto sviluppare l'analisi percettiva ampliando il campione di partecipanti, includendo i 49 soggetti esclusi per via degli audio di produzione disturbati. Questo consentirebbe di confermare i risultati già osservati ma su una base più solida. Sarebbe inoltre interessante avviare uno studio longitudinale riproponendo lo stesso esperimento ai soggetti che l'hanno già svolto in modo da poter confrontare dei dati in *real time*.

Rispetto all'analisi della produzione, è auspicabile proseguire con delle indagini strumentali più accurate che consentano di raccogliere dati acustici e articolatori, eventualmente includendo anche altri contesti fonologici. In questo modo sarebbe certamente possibile delineare un quadro più preciso rispetto alle produzioni degli apprendenti.

Bibliografia

Al Tubuly, S. (2018). *The perception and production of Arabic consonants: A cross-linguistic study*. In Alhawary, M.T. (a cura di), The Routledge Handbook of Arabic Second Language Acquisition, Routledge, New York (ny), 0-92.

Al-Aloula, M. (2018). *Production of Modern Standard Arabic lexical stress cues by native speakers of American English*. In Alhawary, M.T. (a cura di), The Routledge Handbook of Arabic Second Language Acquisition, Routledge, New York (ny), 38-55.

Banfi, E., Grandi, N. (2008). *Lingue extraeuropee: Asia e Africa*, Carocci, Roma.

Best, C.T. (1995). *A direct realist view of cross-language speech perception*. In Strange, W. (a cura di), Speech Perception and Linguistic Experience: Issues in Cross-Language Research, York Press, Timonium (md), 71-206.

Best, C.T., Tyler M.D. (2007). *Nonnative and second-language speech perception: Commonalities and complementarities*, in Munro, M.J.; Bohn, O.S. (a cura di), Language experience in second language speech learning, John Benjamins, Amsterdam, 13-34.

Bloomfield, L. (1933). *Language*, Holt, New York.

Brown, C.A. (1998). *The role of the L1 grammar in the L2 acquisition of segmental structure*, Second Language Research, 14(2), 136-193.

Canepari, L. (2006). *Avviamento alla fonetica*, Einaudi, Torino.

Chini, M. (2005). *Che cos'è la linguistica acquisizionale*, Carocci, Roma.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale.

De Dominicis, A. (2013). *Fonologie comparate. Suoni e lingue d'Europa, Cina e mondo arabo*, Carocci, Roma.

Durand, O. (2009). *Dialettologia araba*. Carocci, Roma.

Eads, A., Khater, J., Mielke, J. (2018). *Arabic L2 phonological acquisition: An ultrasound study of emphatics and gutturals*, in Alhawary, M.T. (a cura di), The Routledge Handbook of Arabic Second Language Acquisition, Routledge, New York (ny), 93-112.

Eckman, F.R. (1977). *Markedness and the contrastive analysis hypothesis,* Language learning, 27(2), 315-330.

Elmahdy, M., Gruhn, R., Minker, W. (2012). *Novel Techniques for Dialectal Arabic Speech Recognition,* Springer, New York.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*, Tesi di dottorato, Utrecht University, LOT dissertation series 113, Utrecht.

Escudero, P., Benders, T., Lipski, S.C. (2009). *Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners*, Journal of Phonetics, 37(4), 452-465.

Fischer, W. (1972). *Grammatik Des Klassischen Arabisch*, Harrassowitz, Wiesbaden.

Fischer, W. (2011). *Classical Arabic*, in Versteegh, K. (a cura di), Encyclopedia of Arabic Language and Linguistics, volume 1, Brill, Leiden.

Flege, J.E. (1995). *Second language speech learning: Theory, Findings, and Problems*, in Strange, W. (a cura di) Speech perception and linguistic experience: Issues in cross-language research, York Press, Timonium (md), 233-277.

Flege, J.E., Bohn, O.S. (2021). *The revised speech learning model (SLM-r)*, in Wayland, R. (a cura di) Second language speech learning: Theoretical and Empirical Progress, Cambridge University Press, Cambridge.

Greenberg, J.H. (1963). *The Languages of Africa*, International Journal of American Linguistics, 29(1).

Honikman, B. (1964). *Articulatory settings*, in Abercrombie, D.; Fry, D.B.; MacCarthy, P.A.D.; Scott, Norman C.; Trim, J.L.M. (a cura di), In Honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday, 12 September 1961, Longmans, Green & Co., London, 73-84.

Kallas, E. (2005). Le sfide didattiche dell'arabo, Oriente Moderno, 85(1), 149-161

Lado, R. (1957). *Linguistics across cultures,* University of Michigan Press, Ann Arbor (mi).

Larcher, P. (2012). *Le système verbal de l'arabe classique*, Publications de l'Université de Provence, Aix-en-Provence.

Mairano, P., Contreras Roa, L., Capliez, M., Bouzon, C., (2021). *The /s/ ~ /z/ voice contrast in L1 French, L1 Spanish and L1 Italian learners of L2 English*, Language, Interaction and Acquisition, 12(2), 210-250.

Manca, A. (1989). *Grammatica (teorico-pratica) di arabo letterario moderno*, Associazione nazionale di amicizia e di cooperazione italo-araba, Roma.

Maturi, P. (2006). *I suoni delle lingue, i suoni dell'italiano: introduzione alla fonetica*, Il mulino, Bologna.

Mion, G. (2010). *Sociofonologia dell'arabo. Dalla ricerca empirica al riconoscimento del parlante*, Edizioni Nuova Cultura, Roma.

Mion, G. (2016). *La lingua araba*, Carocci, Roma.

Mion, G., D'Anna, L. (2021). *Grammatica di arabo standard moderno. Fonologia, morfologia e sintassi*. Hoepli, Milano.

Mori, L. (2007). *Fonetica dell'italiano L2. Un'indagine sperimentale sulla variazione nell'interlingua dei marocchini*, Carocci, Roma.

Newman, D.L. (2002). *The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad*, Antwerp papers in linguistics, 100, 65-75.

Neyreneuf, M., Al-Hakkak, G., (1996). *Grammaire active de l'arabe littéral*, Librairie générale française.

Regueira, X.L. (1996). *Galician*. Journal of the International Phonetic Association, 26(2), 119-122.

Rohlfs, G. (1966). *Grammatica storica della lingua italiana e dei suoi dialetti. Fonetica*, Einaudi, Torino.

Schultz, E. (2004). *Modernes Hocharabisch Grammatik*, Reichert Verlag, Wiesbaden.

Shehata, A. (2018). *Native English speakers perception and production of Arabic consonants*, in Alhawary, M.T. (a cura di), The Routledge Handbook of Arabic Second Language Acquisition, Routledge, New York (ny), 56-69.

Skinner, B.F. (1957). *Verbal behavior,* Appleton-Century-Crofts, New York.

Thelwall, R., Sa'Adeddin, M.A. (1990). *Arabic*, Journal of the International Phonetic Association, 20(2), 37-39.

Tresso, C.M. (1997). *Lingua araba contemporanea. Grammatica ed esercizi*, Hoepli, Milano.

Werker, J.F., Logan, J.S. (1985). *Cross-language evidence for three factors in speech perception*, Perception & Psychophysics, 37(1), 35-44.

Zaltz, Y., Segal, O. (2021). *The Perception Of Arabic Vowel Duration By L1 Hebrew Speakers: Can A Short Training Remold The Effect Of The Native Phonological System?*, Studies in Second Language Acquisition, 44, 143-163.

DOI: 10.17469/O2111AISV000023

BIANCA MARIA DE PAOLIS, ANTONIO ROMANO, CECILIA ANDORNO

Prosodic and syntactic markers of narrow focus in Italian L1 and L2 (French L1): an experimental study

Information structure marking can be a challenge for L2 learners: first, because it implies an articulated interplay of prosody, syntax and word order; second, because the interaction of these linguistic means is subject to language-specific constraints. In this paper, we analyse the narrow focus (non-corrective and corrective) marking strategies of native speakers of Italian and French, and French learners of Italian L2, through a corpus of task-elicited speech (41 speakers). We examine to what extent L2 learners use similar linguistic structures to those observed in the target language in narrow focus marking, and how L1 may influence their productions. Our results show that crosslinguistic effects arise both at the syntactic and the prosodic level, and depending on the contexts, can have a positive or negative effect on learners' productions. While crosslinguistic similarities help in the early mastery of cleft sentences for NF marking, systemic differences at the prosodic level initially prevent prosody from being exploited to the same goal.

Keywords: focus, prosody, syntax, second language acquisition, Italian L2.

1. *Introduction*

1.1 Focus and contrast

Focus is a key-concept in linguistics, and also a hard-to-define one. In a functionalist perspective, focus can be described as the most informative part of the utterance in the speaker's intentions, i. e. the one which conveys the most dynamic element within the communicative act (Halliday, 1967; Vallduví, Engdahl, 1996). In a syntagmatic relation, it is opposed to background, which, conversely, conveys less significant (given, expected, assumed) information (Krifka, 2008). In a situation like the one in example (1b), for instance, the subject "Maria" would be the focused constituent, and "l'ha comprato" ("bought it") would be the background, since the fact that a newspaper was bought is proposed as uncontroversial in both the question and in the answer, and the main contribution for conversation is made by the introduction of the agent, Maria.

(1) (a) Chi ha comprato il giornale?
'*Who bought the newspaper?*'
(b) L'ha comprato [Maria]$_{F}$.
'Maria bought it.'

Focus can be associated with the notion of contrast: some authors argue that the presence of focus itself sets out a selection of contrastive alternatives (Rooth, 1992;

Krifka, 2008); others distinguish neutral-focus types from contrastive-focus types, depending on the overtness or covertness of the contrast carried by focus in relation to these alternatives (Kiss 1998). Mediating between the two views, we can say that the degree of contrast entertained by the focused constituent with its potential alternatives constitutes a continuum, in which identification would be the least contrastive function of focus, and correction the most overt case of contrast (Repp, 2016; Cruschina, 2021). In this way, example (1b) would be a less contrastive (but not completely neutral) type of focus, since the answer "Maria" is not opposed explicitly to its hypothetical alternatives to the candidate focus; example (2b), opposedly, would be a more contrastive one, involving an overt opposition between two explicit alternative agents, Maria and Giulia:

(2) (a) Giulia ha comprato il giornale.
'*Giulia bought the newspaper*'
(b) No, $[\text{Maria}]_F$ l'ha comprato, non Giulia.
'*No, Maria bought it, not Giulia.*'

1.2 Focus and its expression

The articulation between focus and background, as well as the function (identification or correction) of the focused constituent, can be conveyed through different means, involving multiple linguistic levels: morphology, syntax, word order, lexical particles, prosody. With regard to syntax, several means can be exploited, to an extent that depends on the discursive situation and language-specific constraints: postverbal subject, focus fronting, dislocations, etc. (Lambrecht, 1994; Neeleman, van de Koot, 2014). Cross-linguistically, a structure that is frequently associated with focus is the it-cleft sentence (Dufter, 2009; De Cesare, 2017), in which focus is expressed by the clefted constituent and the relative clause contains the background, as in (3):

(3) È $[\text{Maria}]_F$ che ha comprato il giornale.
'It's Maria who bought the newspaper.'

The use of clefts to express focus-background articulation is largely attested in both Germanic and Romance languages; its interpretation in terms of exhaustiveness, identification or correction, though, is not equivalent cross-linguistically (De Cesare, Garassino, 2015). Some languages specialize this construction for more contrastive situations (e.g. English or Italian), others use them in a larger array of contexts (e.g. French); it has been argued that this divergence is due to differences in the language phonological structure and word-order flexibility (Lambrecht, 2001; Destruel, DeVeaugh-Geiss, 2018; Greif, Skopeteas, 2021; Larrivée, 2022).

Besides (and along with) syntax, prosody plays a key role in the expression of information structure. Its function lies in highlighting the focus constituent, making it stand out from the rest of the utterance (Selkirk, 1984; Baumann, 2005; Baumann, Riester, 2012); the way in which prominence is implemented, though, is subject to great variability. It is often said that head-prominence languages, such as German or English, mark focus through pith-accents (Cruttenden, 2006; Büring,

2016), whereas edge-prominence languages, like for example Japanese or French, mark it by phrasing (Jun, 2005; Féry, 2013). Recent work, however, has highlighted a more complex scenario: a better definition would be that of a continuum, spanning from head-prominent to edge-prominent type, along which languages are placed according to their employment of different strategies (Sbranna, Ventura, Albert & Grice, 2023). Furthermore, it has been observed that languages can share the same syntactic and prosodic means to encode focus, but use them in different pragmatic circumstances (Vander Klok, Goad & Wagner, 2018).

1.3 Information structure marking and SLA

In recent years, there has been an increasing interest in the study of the syntax-prosody interplay and its implications for Second Language Acquisition. Prosody alone represents a problematic area for L2 learners: it's an important factor in the perception of "foreign accent" (Boula de Mareuil, Vieru-Dimulescu, 2006), and an area in which L1 transfer seems to be particularly persistent (Mennen, 2004). In the expression of information structure, the difficulties of prosody sum up to those of its mapping with syntax[1]: this combination often represents a "final hurdle" learners never overcome, not even in advanced stages of acquisition (Mennen, 2015). Traditional models of L2 acquisition have predominantly focused on phonology and phonetics at the segmental level, as evident in the works of Flege (1995) and Best (1995), while substantially neglecting suprasegmental traits such as rhythm, intonation, and phrasing. Noteworthy attempts have been made in recent times to address this gap (So, Best, 2008; Gili Fivela, 2012; Wiener, Goss, 2018); nevertheless, these endeavors have yet to culminate in a comprehensive model. Prosodic aspects still represent areas under exploration, yielding mixed results and unresolved issues (Mennen, de Leeuw, 2014; Gili Fivela, Avesani, 2021). If several studies suggest that learners inevitably remain sensitive to the constraints of their L1 in expressing information structure (Hertel, 2003 and Lozano, 2006 for L2 Spanish; Zubizarreta, Nava, 2011 for L2 English), there is no consensus as to whether L1 transfer is active at the level of grammar or phonological categories, or at the level of phonetic detail, realisational dimension, and "nativelike" preferences.

Another point that has not yet been clarified is the role that typological proximity between source and target language plays in the acquisition of information-structural markers (Benazzo, Andorno, 2017). In fact, it has been argued that the effects of cross-linguistic influence (CLI) may be different depending on the typological proximity of the two languages in question, and that different combinations lead to different results; all the more so in the relationship between syntax and prosody (Andorno, Turco, 2015; Benazzo, Dimroth & Santiago, 2021).

[1] We acknowledge the existence of formal models that conceptualize the matter under consideration as an "interface" concern, as advanced by Sorace (2006). Nevertheless, for the purpose of this paper, we choose to adopt a functionalist perspective, deliberately eschewing a formalist lens.

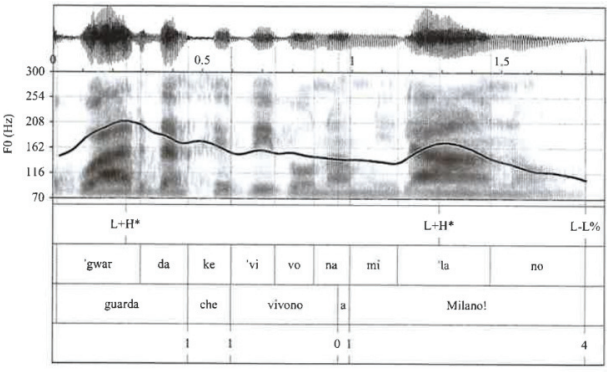
2. *Background and research goals*

2.1 Expressing focus in Italian and French

In this work we are interested in the description of two cognate languages, French and Italian, and the influence of L1 French on Italian L2 learners' production. The first difficulty we run into during our investigation is the description of the two languages in relation to their prosodic type (§ 1.1): are French and Italian head-prominent or edge-prominent languages? Moreover, if it is quite straightforward to say that Italian has a lexical accent, what can we say about accent - and consequently pitch-accents location - in French (Cole, Hualde, Smith, Eager, Mahrt & de Souza, 2019)?

Italian is a stress language, with penultimate stress being the most common pattern, and it has traditionally been classified as a syllable-timed language (Bertinetto, 1981). In Italian, and especially in the variety of Turin, narrow focus is realised as an L+H* L- pitch accent (Avesani, 2003; Avesani, Vayra, 2004; Gili Fivela, Avesani, Barone, Bocci, Crocco, D'Imperio, Giordano, Marotta, Savino & Sorianello, 2015). Prominent constituents are also signaled by the use of wider pitch range, longer duration and higher intensity of the accented syllable (Avesani, 2003). An example of the f0 contour for a narrow-contrastive focus statement is shown in Fig. 1.

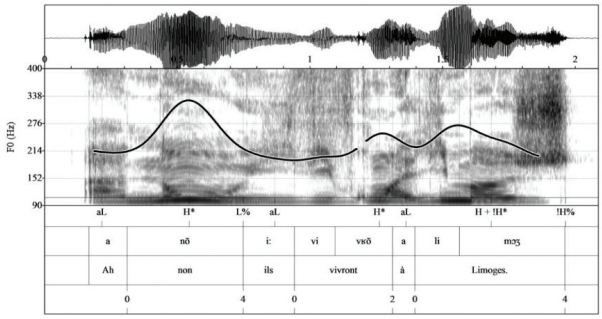
Figure 1 - *Contradiction statement "No, guarda che vivono a Milano" ("No, they live in Milan") uttered by an Italian speaker from Florence (from Gili Fivela et al., 2015)*



French, on the other hand, has a final accent, whose function is not lexical, but edge-marking on the right boundary of the accentual phrase (AP); apart from this final accent, known as obligatory, it is also possible to place an accent initial (Hi) to highlight a focused constituent (Astésano, Bertrand, 2016; Di Cristo, 2016; Michelas, German, 2016). The focus-accent association is not limited to metrical patterns: in addition to the initial accent, a focused AP can be produced with a significant final descending melodic contour, tonal enhancement and slower speech rate. All these features convey a contrast between focused and unfocused APs, the latter produced without Hi (initial) accent and with a final rise (Hirst, Di Cristo, 1996; Jun, Fougeron, 1995, 2000; Delais-Roussarie, Post, Avanzi, Buthke, Di Cristo,

Feldhausen, Jun, Martin, Meisenburg, Rialland, Sichel-Bazin & Yoo, 2015), that could be described as a continuation contour (Delattre, 1966). Figure 2 shows an example of f0 contour for a narrow-contrastive focus statement in French, featuring an initial accent followed by a steep f0 fall on the focus word "Limoges".

Figure 2 - *Contradiction statement "Ah non, ils vivront à Limoges" ("No, they will live in Limoges"), uttered by a native French speaker from Marseille (from Delais-Roussarie et al. (2015)*



As we have stated in the introduction (§ 1.3), mastering the interplay between prosody and syntax in the expression of information structure is a challenge for L2 learners. A few studies have addressed the question with regards to Italian L2, confirming the persistence of L1 influence in this area, although with different outcomes – mostly depending on the distance between source and target language. In Turco, Braun & Dimroth (2015), West-Germanic speakers are said to show effects of CLI in the use of pitch accents and verum focus to mark polarity focus in Italian. In Avesani, Bocci, Vayra & Zappoli (2015) German learners appear to have acquired the (negative) relation between discourse properties and their prosodic marking that holds in Italian, but show signs of transfer on the phonological level.

To our knowledge, the question has never been addressed taking into account a combination of close-related languages, as would be the case with French learners of L2 Italian. We think this track could lead to interesting and, especially, complementary results from what we have already observed with Germanic source language learners.

2.2 Research questions

Our goal in this work is to observe the productions of non-native speakers, and try to interpret them in terms of cross-linguistic influence and L2 acquisition universals. The novelty of our study lies not only in its consideration of a little-studied combination of languages, but also in combining the syntactic and prosodic perspectives. In this context, we aim to assess the potential contributions of a methodology explicitly designed to harmonize these two dimensions; notably, our approach involves the analysis of both prosody and syntax in semi-spontaneous speech. While we are aware that analysing prosody through non-controlled material is not straightforward, we

think that fully-scripted speech would reduce the validity of conclusions, especially in investigating an information-structural category such as focus.

3. *Methodology*

3.1 Population of the study

As already stated, our study aims at observing and describing the production of Italian L2 speakers, and comparing them to the production of native speakers of both the source and the recipient languages, i.e. Italian and French. For this purpose, we chose to set up a multi-group, interindividual design, and recruited 3 groups of participants: the test group, made up of native French speakers learning Italian as L2, along with two control groups, made up of native French (with no or little competence in Italian) and native Italian speakers (with no or little competence in French). Speakers of the learner group acquire their L2 through non-guided learning (in the terms of Klein, Perdue, 1997; see also Klein, Dimroth, 2009): they are adult workers or students living in Italy, not attending any Italian language courses.

Participants were recruited and recorded between summer 2021 and fall 2022, in Turin, Italy (Italian L1 and Italian L2 groups) and Paris, France (French L1 group). Special care was taken in circumscribing the speakers' areas of origin, to minimize the impact of diatopic variation. For Italian, we adopted as a point of inquiry the area of Turin, whose variety is described, with particular interest for prosodic focus, in Romano, Interlandi (2008). The composition of the groups is shown in Tab. 1.

Table 1 - *Population sample of the study*

Group	*Proficiency*	*N. of speakers*	*Region of origin*
ITL1	-	14	Torino
FRL1	-	14	Paris
	B1 / B2	7	
ITL2	C1 / C2	6	Paris

Proficiency levels were assessed through three complementary tests: self-assessment, written cloze test (Vedder, 2008), evaluation of oral productions by L2 Italian teachers in accordance with CEFR guidelines (Council of Europe, 2001)[2].

[2] Since the corpus was constructed for a study on acquisitional phenomena, we took care to assess the speakers' proficiency level, which is why we report it in the sample description. Considerations about this factor remain, however, beyond the scope of the present article, in which we want to concentrate on observing the variety of marking strategies in the L1 and L2 groups, and not their relation to speaker' proficiency levels. In any case, level of competence seems not to be a dominant parameter for the analyses here reported. For an in-depth inquiry in that direction, we refer to Andorno, De Paolis (submitted).

3.2 Data collection protocol

Experimental studies on information-structure bring up an important methodological issue: on one hand, one has the need to analyse speech that conserves (at least partial) ecological, conversational value; on the other hand, the perspective of a phonetic analysis asks for the elicitation of controlled, fully comparable prosodic units. These two requirements are obviously contradictory and pose problems for designing the collection protocol. Moreover, one must take into account the need to compare results with those from other language combinations, an essential step in the identification of CLI. In an attempt to meet all these requirements, we decided to adopt the protocol designed by Gabriel (2010) and later taken up by Feldhausen and Vanrell Bosch (2016) in similar research, which is based on task-elicited speech. Stimuli for the protocol were translated into the languages useful for our purpose, Italian and French. The task is carried out as follows: first, the participant is shown PowerPoint slides containing two short stories, each one accompanied by a caption; then, the participant is shown other slides with the same pictures, accompanied by questions (see Fig. 3 in the next page). The participant is asked to answer each question aloud in complete sentences[3], with no other indications on how to structure the utterance other than a request to avoid one-word answers.

Figure 3 - *Stimuli from the Italian version of the task*



[3] In the prompt, the term "complete sentence" was intended to refer to sentences including a finite verb. This specification aimed to prevent elliptical answers such as "Maria.", which would not be conducive to the study of syntactic marking and word-order changes. Regrettably, there was an unintended interpretation of this prompt, where participants understood it as an obligation to produce "canonical sentences", leading to an avoidance of non-SVO orders. Consequently, in comparison to other corpus studies, our dataset exhibits lower occurrences of post-verbal subjects, dislocations, fronting, and other structures deviating from the canonical SVO order.

(4) (a) Che cosa succede qui?
'What's going on here?'
(b) Chi compra il giornale in edicola?
'Who's buying the newspaper at the newsstand?'
(c) Giulia dà il giornale a suo fratello, giusto?
'Giulia is giving the newspaper to her brother, right?'

Questions are formulated to elicit three types of focus-utterances: (a) broad focus, (b) narrow identification focus, (c) narrow correction focus. As shown in the examples in (4), identification focus is elicited through a partial question, and corrective focus through an assertive statement accompanied by a tag-question[4], through which the reaction (in this case, corrective) of the speaker is sought. Partial questions and tag-questions target different syntactic components of the sentence: (i) subjects, (ii) verbs, (iii) direct objects, (iv) indirect objects, (v) adverbials. Participants are alone in the recording booth and can advance with the slides at their own pace. Recordings were carried out in a soundproof environment, with a Focusrite Scarlett interface and Shure SM58 microphones. Audio files were recorded in .wav format, with a sampling rate of 44100 Hz.

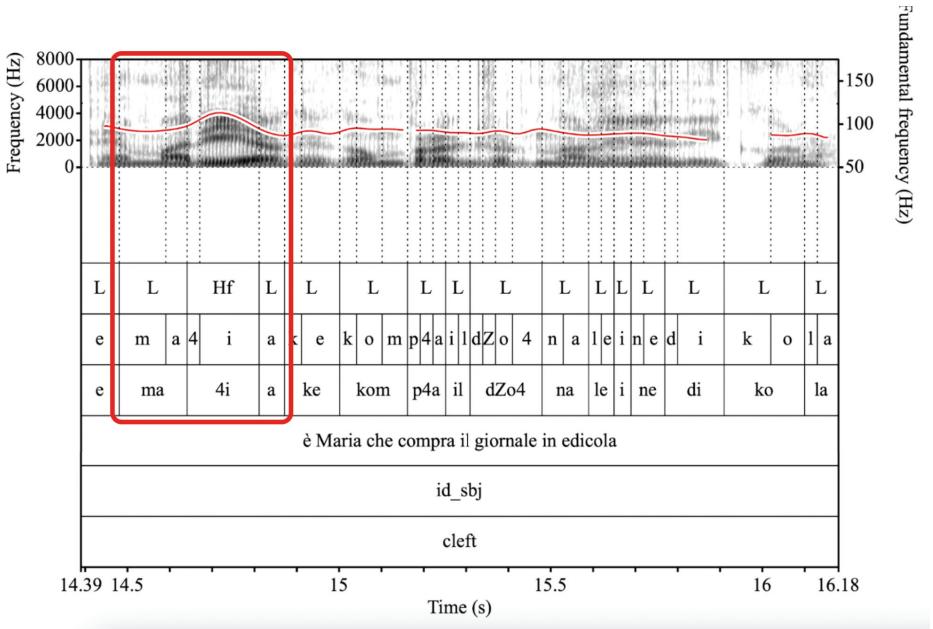
3.3 Tools for the analysis

Transcription and labeling of syntactic phenomena was done on the basis of a taxonomy taken, with minor adaptations, from Brunetti, Bott, Costa & Vallduví (2011). In our corpus, it-cleft sentences proved to be preponderant as a marked structure in both languages (for a precise account of their frequency, we refer to § 4). For simplicity, therefore, we decided to restrict our analysis to a sub-corpus including either it-cleft sentences or unmarked sentences. Instances of other non-canonical orders (e.g., postverbal subjects, clitic dislocations) were excluded. Segmentation at the word, syllable, and phone level was done via WebMaus (Kisler, Reichel & Schiel, 2012). A not insignificant problem arose with regard to prosodic labeling. If, in fact, if ToBI – in ToBIT and FrToBI versions, see Avesani (1995) and Delais-Roussarie (2015) – has been used in many works on L1 intonation, it poses a major problem for L2 studies. ToBI, in being a phonological type of annotation, presupposes the existence of a categorical and grammatical representation of the functions attributed to suprasegmental articulation in the speaker's system. This assumption, however, does not hold in the case of L2 speakers: a speaker's interlanguage, in fact, is a system of its own that cannot be assumed to be identical to that of either its source or target language (Selinker, 1972; Eckman, 1981; Mack, 1988). We chose, therefore, to code intonation through the Polytonia script (Mertens, 2014), which processes prosodic labels from raw phonetic values, without phonological interpretation, in relation to the individual characteristics of each speaker (pitch range, speech rate, etc.). The algorithm provides a series of symbols that represent tonal levels and the perceptible melodic movements of each syllable, based on the notion of the seuil de glissando

[4] On the role of tag questions as biased questions see Romero (2020).

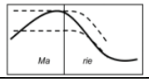
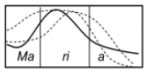
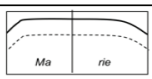
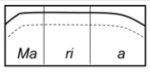
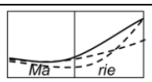
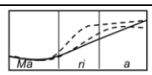
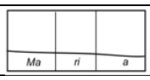
(Rossi, 1971). The system provides 5 symbols for the tonal level: L (low), M (mid), H (high), B (bottom) and T (top). It also provides 5 melodic movements: R (large rise), F (large fall), r (small rise), f (small fall), and _ (level). Finally, the system can provide for each syllable a combination of these symbols. Fig. 4 shows an example of Polytonia output, together with the other tiers of transcription and labeling.

Figure 4 - *Example of Polytonia labeling for identification-focus statement "È Maria che compra il giornale in edicola" ("It's Mary who's buying the newspaper at the newsstand") by a male Italian speaker*



As can easily be imagined, a range of 5 levels plus 5 movements gives rise to a large number of combinations; the possible sequences on constituents of 2 to 4 syllables thus become numerous. For this reason, we decided to re-elaborate the taxonomy, narrowing down from the combinations provided by Polytonia to 4 macro-types, established on a functional basis (for details we refer to De Paolis, Santiago & Andorno, 2022). This allowed us to reduce the fragmentation of the results and identify sharper trends. The rationale and outcome of this reclassification is shown in Tab. 3. The intonation of target words, such as "Maria" or "giornale", was profiled by combining the Polytonia labels assigned to individual syllables. For instance, determining whether "Maria" exhibited a falling, flathigh, flatlow, or rising pattern involved assembling the Polytonia labels associated with its three constituent syllables ([ma.ˈriː.a]) to ascertain the overall prosodic type of the word. An example of a target word featuring a falling contour is visible in the previous page, in Fig. 4.

Table 3 - *Prosodic types retained from Polytonia, with the example Marie/Maria (from De Paolis et al., 2022, our translation)*

Prosodic type	Language	Polytonia labels	Stylization	Interpretation
falling	FRL1/L2	MR-HF, M-Mf, r-f, r-HF		A rise / high tone is perceived at the beginning of the first syllable (Hi), followed by an important melodic fall on the last syllable, showing a major break.
	ITL1	Lr-HF-L, 0-LR-HF, M-H-L, 0-H-L,		A melodic rise/high tone is perceived in the syllable carrying the lexical accent, followed by an important fall on the last syllable, showing a major break.
flathigh	FRL1/L2	M-M, H-H,		The entire constituent is perceived in the high or medium range in the speaker's pitch range, indicating a tonal enhancement.
	ITL1	H-H-H, 0-H-H, 0-M-M, M-M-M,		
rising	FRL1/L2	L-LR, M-H, L-M, L-Mr, L-MR,		A rise is perceived on the last syllable of the constituent, indicating the boundary of the accent phrase in the non-final position of the utterance (continuation rise).
	ITL1	L-0-H, 0-L-M, L-Lr-M		A rise is perceived in the tonic syllable of the word, followed by a rising contour (or high plateau) in the postonic (continuation rise).
flatlow	FRL1	0-0, L-L, 0-L		The entire constituent is perceived in the speaker's low pitch range, no melodic movement is perceived.
	ITL1	0-0-0, L-L-L, 0-0-L, 0-L-0		

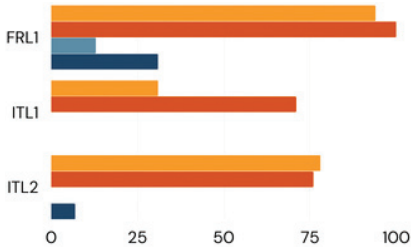
4. *Results*

Our dataset consists of 615 utterances in total: 15 per speaker, for 41 speakers (14 ITL1, 14 FRL1 and 13 ITL2). In this study, we observe the realisation of subjects and direct objects. We will illustrate the results of this analysis in two stages: first, with a focus on syntactic marking; second, with one on intonation; some considerations about the combination of the two will be made in § 5.

4.1 Syntactic marking

The first step of our analysis was to observe the use (number and context of occurrences) of it-clefts by speakers of the three groups. The proportion of such structures realised by speakers of each group in the two narrow-focus contexts (identification and correction) and for each syntactic role (subject and object) is shown in Fig. 5.

Figure 5 - *Proportion of cleft sentences realised by speakers of each group for identification-subjects (yellow), correction-subjects(orange), identification-objects (light blue), correction-objects (blue)*



In all three groups, both contrast and syntactic role significantly influence the use of it-clefts: the subject role and the corrective function enhance the likelihood of clefting. This pattern can be attributed to the specific communicative demands of subjects: as they naturally occupy the initial position in an utterance – a position typically reserved for topics, according to communicative principles (Givón, 1989) – they require additional marking to align with this function. That said, interesting differences can be observed among the three groups. French speakers make large use of it-clefts in both narrow-focus conditions, making it difficult to observe an effect of contrast: in the subject-focus context, the percentage of utterances marked by this structure is very high for both identification (94 %) and correction (100 %). A considerably lower number of occurrences is observed for the object-focus marking: 14 % for identification and 31 % for correction. It is interesting to notice, anyway, that the difference between identification and correction is in both cases (subject vs object) not significant ($p>0.05$ for both subjects and objects). As for the native Italian group, there is no occurrence of it-cleft structures for object marking, the use of this structure being limited to subjects. Regarding the role of contrast, it-clefts can be used for identification focus, but do not represent the default strategy (31 % of cases). Their frequency increases for correction focus (71 %); still, without reaching the proportion of the French group. Anyway, the difference between the two conditions, identification vs correction, is for the Italian group highly significant ($p<0.0005$).

For Italian L2 speakers, the results indicate a "middle ground" between the two native groups. The data reveal that French learners produce fewer it-cleft sentences overall compared to the FRL1 group, indicating partial approximation to the target language. However, they do not appear to distinguish between the two conditions – identification and correction – as evidenced by the similar percentages of clefted utterances: 78 % for subject identification and 76 % for subject correction ($p > 0.05$). For object marking, slightly more cleft sentences were produced in the corrective context, but the difference remains statistically insignificant ($p > 0.05$). The interaction between syntactic role and focus type thus differs in French and Italian: in French, focused subjects consistently trigger it-cleft constructions, regardless of sentence function; in Italian, syntactic role

and focus type both have an impact on cleft frequency. To test this interpretation of our data, we applied a linear regression model with two-factor interaction using the lm package in R (R Core Team, 2021):

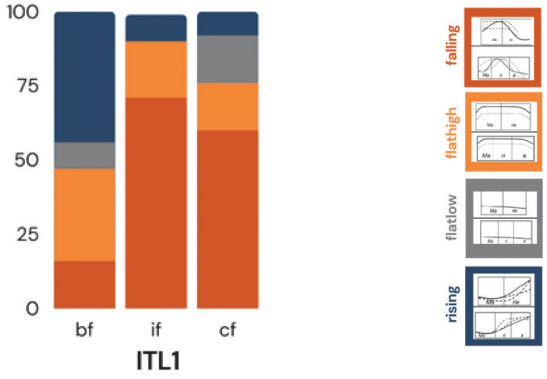
*model <- lm(cleft ~ infContext * tgtConst)*

For Italian, the model exhibits a robust fit (Multiple R-squared: 0.7082, Adjusted R-squared: 0.6936, F-statistic: 48.53 on 3 and 60 DF, p-value < 0.0005). In contrast, the model lacks effectiveness for French, where the sole reliable predictor is the syntactic role. Regarding ITL2, the model suggests a modest fit (Multiple R-squared: 0.2631, Adjusted R-squared: 0.2236, F-statistic: 6.664 on 3 and 56 DF, p-value < 0.0005).

4.2 Prosodic marking

The prosodic analysis was performed on the same dataset, including all utterances that were either unmarked (i.e., exhibiting SVO order without dislocations, clitic anticipation, or clitic resumption) or marked using it-clefts[5]. In this section we will show for each of the three groups the results of a quantitative analysis conducted on the basis of Polytonia output. Fig. 6 summarizes the proportion of prosodic types observed in the native Italian group, according to the focus subtype, broad vs identification vs correction.

Figure 6 - *Distribution of contours in broad, identification, correction focus utterances (group ITL1)*



[5] We chose to incorporate cleft sentences alongside those with a canonical order in the same dataset for two primary reasons. Firstly, literature descriptions suggest that the prosodic profile of clefted constituents does not structurally differ from that of in-situ focus constituents (Mertens, 2012; Pinelli, Poletto, Avesani, 2020). Secondly, our aim was to analyse syntax and prosody as additive strategies rather than complementary ones (a point we will revisit in the discussion, see par. 6). Therefore, it would not have been meaningful to solely describe clefted constituents from a syntactic perspective and in-situ constituents solely from a prosodic one. It's also worth noting that separating clefted from non-clefted constituents would have created an imbalance between the French and Italian datasets, due to the notable frequency disparity in cleft structures observed in the two groups.

As the chart shows, in the broad focus (bf) condition the vast majority of constituents "Maria" and "giornale" are realised with a rising continuation contour (44 %) or with a general tonal enhancement (flathigh, 31 %). A few realisations present a falling or flatlow contour; this small number represents some realisations of the word "giornale", produced sporadically with this intonational profile because of its quasi-final position ("Maria compra il giornale in edicola", "Maria is buying the newspaper at the newsstand"). When we observe the identification-focus condition (if), we observe a drastic increase in realisations of falling contours (71 %), at the expense of types rising (9 %) and flatlow (0). In contrast, the transition to the corrective-focus condition (cf) is smoother: rising contours still account for a very small portion of the total (8 %), and falling and flathigh contours represent the majority (60 % and 16 % respectively). That 16 % of flatlow contours is made up of realisations as "No, Maria non compra delle parole crociate; Maria compra il giornale in edicola" ("No, Maria isn't buying crosswords at the newsstand; Maria's buying the newspaper at the newsstand"), in which "il giornale", due to its utterance-final position and the explicit negation of the preceding utterance, is realised prosodically as a post-focal, non-prominent, constituent.

Figure 7 - *Prosodic marking of broad, identification, correction focus for group FRL1*

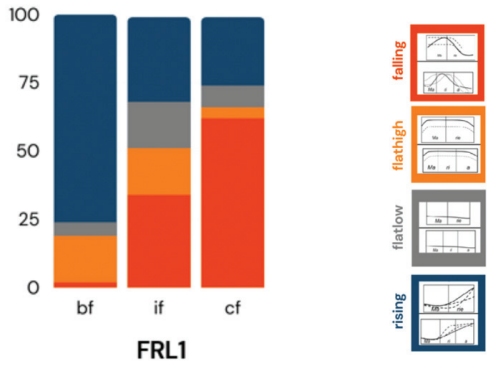
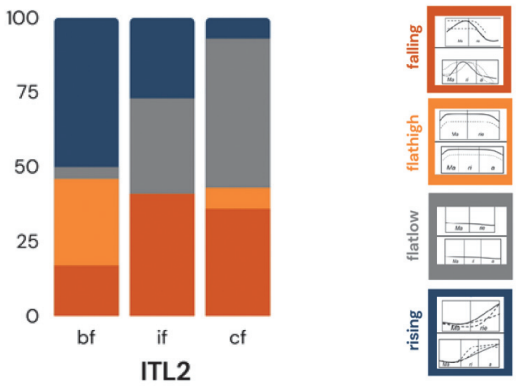


Fig. 7 summarises the proportion of prosodic types observed in the native French group. Similarly to the native Italian group, French speakers realise the constituents "Marie" and "journal" with a large majority of rising (76 %) and flathigh (17 %) contours in the bf condition; falling and flatlow account for a small proportion of the total, 2 % and 5 % respectively. The same considerations made about these two contours in the Italian group hold in this case. Otherwise, a strong difference emerges in the transition from the bf condition to the if condition: French speakers continue to produce a substantial amount of rising contours (31 %), along with the falling and flathigh (34 % and 17 %). Even more strikingly, we find 17 % of flatlow realisations for focused constituents. A similar situation is found for corrective focus: even though the amount of falling

contours rises to 62 %, rising and flatlow realisations still account for 8 % and 25 % of the total. This can be partially explained by resorting to observations made about syntax (see § 3.1): constituents that are not prosodically marked, in fact, often coincide with those that are marked through a cleft-sentence, which, as we saw in the previous section, is massively exploited by French speakers in both contexts of narrow focus.

Figure 8 - *Prosodic marking of broad, identification, correction focus for group ITL2*



When moving to the Italian L2 group, we notice some similarities with the two groups ITL1 and FRL1, and, most interestingly, some crucial differences. The results for the non-native group are shown in Fig. 8. In the bf condition, L2 speakers' behaviour mirrors that of the target group: the large majority of constituents is produced with flathigh or rising contours (29 % and 50 %), while falling (17 %) and flatlow (4 %) account for a small part of the total. A somewhat unexpected situation emerges, instead, when we move to the narrow-focus conditions. In fact, as we can see from the chart, if a rise in the proportion of falling contours is visible (32 %, against 17 % for bf) in the identification-focus (if) condition, a relatively large part of the constituents is realised with a flatlow contour. Moreover, rising contours still account for 27 % of the total. Moving to the most contrastive situation, i.e. correction focus, the proportion of flatlow contours strikingly rises to 50 %, while falling contours drop to 36 %. If, for the FRL1 group, the presence of flatlow contours in contrastive situations could be explained by speakers resorting to syntax, it is not the case for the ITL2 group: in fact, the use of cleft sentences, while more frequent than that made by the ITL1 group, does not fully compensate for this absence of prosodic marking.

5. *Discussion*

5.1 Syntax and prosody in the L1 groups

For what concerns syntax, it emerges quite clearly that it-cleft sentences represent a widely exploited solution in both French and Italian, albeit with important differences between the two languages. As for the overall frequency, our results align with the literature: French speakers produce it-clefts more frequently than Italian speakers, who often choose in-situ marking instead (Lambrecht, 2001; Dufter, 2009). In terms of syntactic role, the tendency is similar in the two groups: it-clefts are used more often to mark focused subjects rather than focused objects. However, while for French this implies only a difference in frequency (94 % and 100 % clefts for subjects in if and cf conditions, and 14 % and 31 % for objects), for Italian there seems to exist a stronger constraint, completely preventing the use of cleft structures for object marking. This kind of asymmetry in subject-object marking has already been attested and accounted for in numerous studies conducted on Italian and French (Belletti, 2015, Samo, Merlo, 2021). Regarding the other variable, however, it seems that the degree of contrastiveness carried by the focus constituent has an effect only in Italian. In fact, while French speakers employ clefts to a similar extent for identification and correction, Italian speakers seem to "specialise" cleft structures to mark correction. For what concerns prosodic marking, results from the native groups confirm the descriptions in literature (§ 2.1): speakers from both groups mark narrow focus through the use of specific f0 contours, with some structural differences related to the phonological features of the two languages, such as the presence of Hi for French, and the anchoring to the accented syllable for Italian (see Fig. 9 and 10). The two charts in Fig. 11 show an important aspect of the two native languages: in Italian, prosody can be used without syntax (orange portion, in-situ focus); the opposite, instead, is not true. In French, the situation is reversed: syntax can be used in absence of prosodic marking (the blue space, representing syntax alone, occupies more than a quarter of the total); while the opposite case, i.e. in-situ focus, is rarely attested. On the other hand, we can observe a common tendency in the L1 groups: in almost half of the cases, the focused constituent is marked by both syntax and prosody. These results also confirm the descriptions of native systems present in the literature (Féry, 2001; Mertens, 2012).

Figure 9 - *F0 contour for the word "Marie" in the identification-focus condition (native French speaker)*

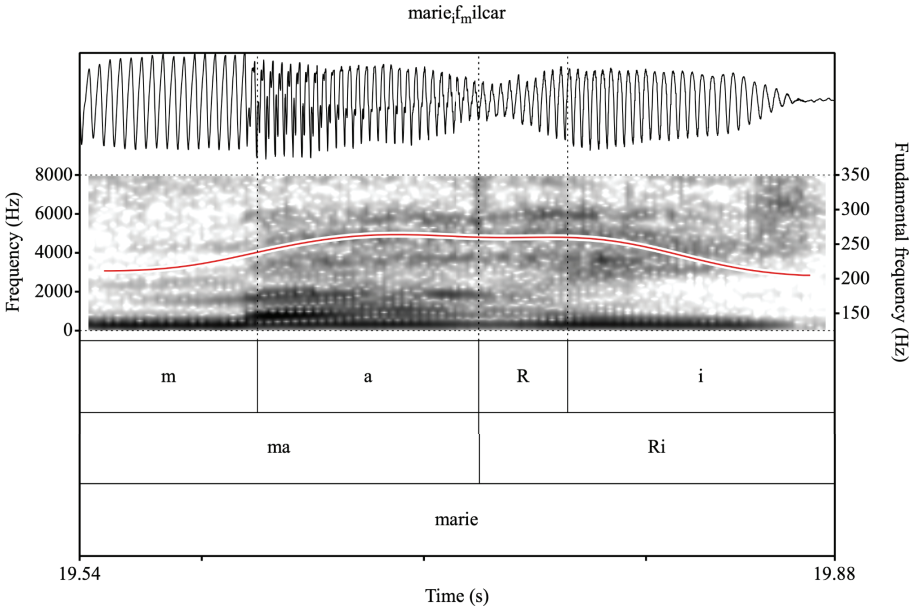


Figure 10 - *F0 contour for the word "Maria" in the identification-focus condition (native Italian speaker)*

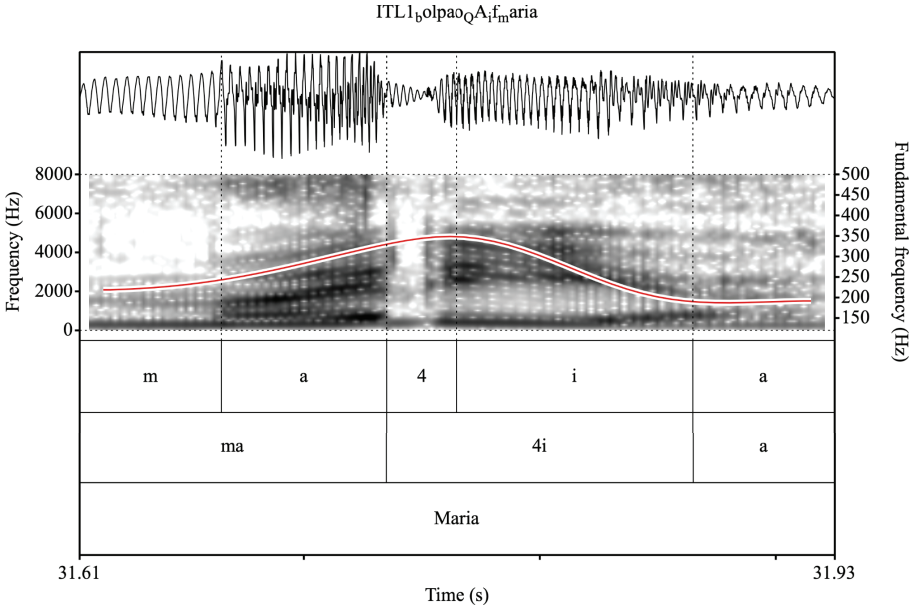
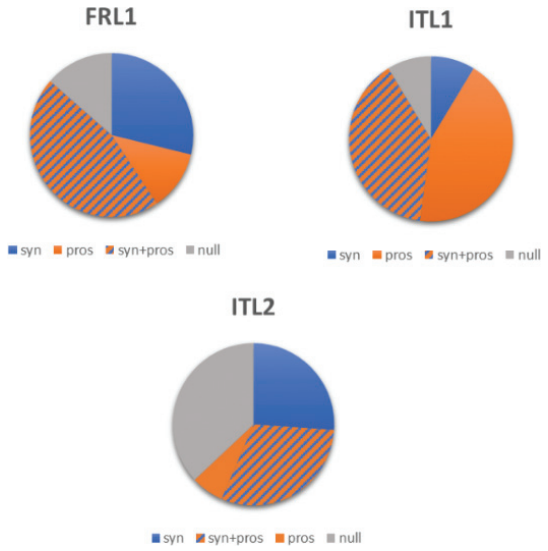


Figure 11 - *Proportion among different focus marking strategies in groups FRL1 (left) and ITL1 (right)*

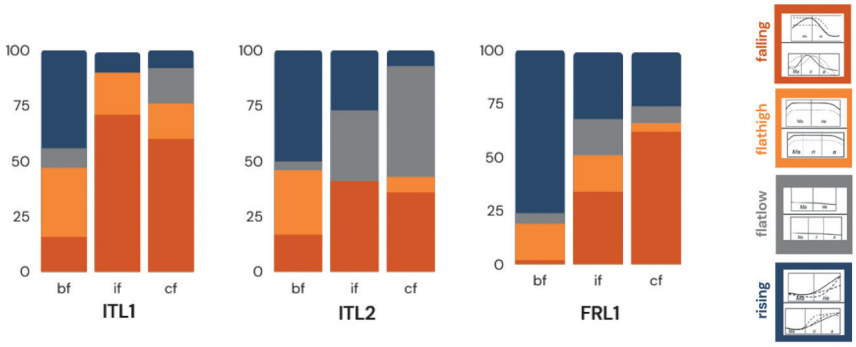


5.2 Syntax, prosody and CLI in the L2 group

The outcome of these differences and similarities between the two L1s can be easily traced in the performance of Italian L2 learners. In fact, despite the potential difficulty in acquiring a complex construction such as the cleft sentence, learners effectively use this structure in the appropriate contexts. Yet, they do it in a way that mirrors the behaviour of their source language, both in terms of absolute frequency and of information-structural context.

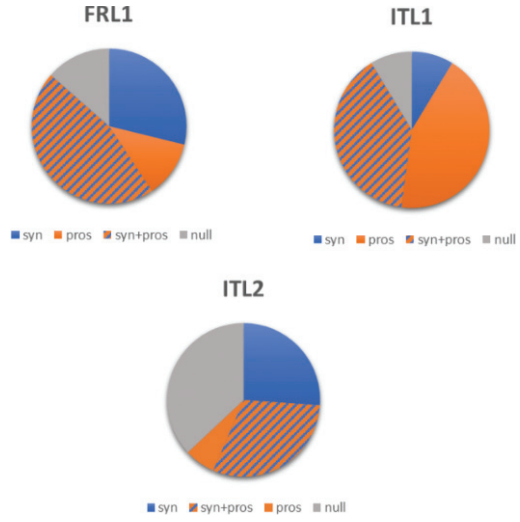
A more complex picture arises, instead, from the analysis of prosodic means used by learners, since they are not the same as those observed in neither of the two L1 groups. In many cases, non-native speakers realise focus phrases with a non-prominent intonation pattern, rising or flat. Constituents' profile seems to be dictated more by their position within the utterance than by their function: a continuation contour is observed if the constituent is at the beginning of the utterance, a flat, low contour if the constituent is at the end. This way, bf and if-cf utterances often share the same prosodic pattern. We propose a comparative visualization of the results of the three groups from the prosodic point of view in Fig. 12.

Figure 12 - *Proportion of different contours (falling, flathigh, flatlow, rising) produced by each group in 3 focus conditions: bf, if, cf.*



A possible explanation for this divergence from the target is that overall disfluency and conflict between different phonological constraints in L1 and L2 inhibit the use of prosodic marking, causing the frequency of prosodically marked structures to be much lower than in the L1 groups (see Avesani *et al.*, 2015). This lack of prosodic marking, not compensated for by a larger incidence of syntactic one, leads to less marking overall, as we can see in Fig. 13 (compared to Fig. 11).

Figure 13 - *Proportion among different focus marking strategies in the non-native group (ITL2)*



These results suggest that, in expressing the information structure of utterance, learners do not rely on prosody to the same extent as native speakers do: in the L2 group, focus/background articulation is realised mostly through syntax. Since this trait cannot be traced back to the influence of L1, one could think of it as a trait

typical of learners' interlanguage. It would not be the first case that such a finding emerges: studies on other language pairs suggest that syntactic marking is more salient for learners than prosodic marking (Yan, Warren, Calhoun, 2022).

6. *Conclusions and perspectives*

In this work we have tried to draw a general picture of the syntactic and prosodic marking means used by Italian, French and Italian L2 speakers to mark narrow-identification and narrow-correction focus. The results show different outcomes of the acquisition process of French learners of Italian L2, depending on the aspect taken into consideration (syntax vs prosody). A possible interpretation for these different outcomes lies in the divergent effects that CLI can have on these two levels, syntax and prosody, depending on the points of contact and of distance of the two native systems from these points of view. Regarding the prosodic analysis conducted on the L2 group, though, an important question arises: can we really say that there is no prosodic marking in absence of major f0 movements[6]? And, also, could it be the case that rising contours are used by learners to mark narrow focus, differently from what observed in the native groups – where rising patterns are usually used for continuation in non-focus phrases? As we have stated throughout the paper, it cannot be assumed that learners' interlanguage will work on the same categories and constraints as the learners' native or target system. A partial attempt to address this question is represented by the work from De Paolis, Abbà, Romano (2023), in which complementary features are considered, in particular the presence of glottal stops as an index for different phrasing patterns.

A last – but still fundamental – question that is yet to-be addressed concerns the role of perception in the acquisition of means for information structure marking. If learners "fail" to make appropriate use of syntax and prosody to highlight focus constituents, can this be attributed to a non-recognition of their functions at the perceptual level? In light of the recognized significance of the perception-production link in an acquisitional context, the next phase in this investigation will involve the development of a perception test to integrate production data.

Acknowledgments

We thank Bianca Abbà for her contribution to data curation, Ilinca Cojan for creating the graphics used as stimuli, Elena Laura Baratono for the evaluation of proficiency L2 speakers, and all the speakers who volunteered for the recordings. This research project is co-funded by the Italian-French University through Projet Vinci 2022.

[6] A similar question has also been addressed by Bocci, Avesani (2011).

References

Andorno, C., De Paolis, B.M. (submitted). Now you see it, now you don't: crosslinguistic effect of syntactic similarities and prosodic differences between similar languages in information structure marking, Paper presented at *ICL2 - Intercomprehension-based L2 learning and teaching*, December 1-2, 2022.

Andorno, C., Turco, G. (2015). Embedding additive particles in the sentence information structure: How L2 learners find their way through positional and prosodic patterns. In *Linguistik Online*, 71(2).

Astésano, C., Bertrand, R. (2016). Accentuation et niveaux de constituance en français: enjeux phonologiques et psycholinguistiques. In *Langue Française,* 3, 191, 11-30.

Avesani, C. (1995). ToBit. Un sistema di trascrizione per l'intonazione italiana. In *Atti delle V giornate di studio del Gruppo di Fonetica Sperimentale,* Trento 1994, 85-98.

Avesani, C. (2003). La prosodia del focus contrastivo. Un accento particolare? In Marotta, G. & Nocchi, N., (a cura di), *La coarticolazione. Atti delle XII Giornate di studio del Gruppo di Fonetica Sperimentale*, Pisa, 28-30 novembre 2002, 157-167.

Avesani, C., Bocci, G., Vayra, M., Zappoli, A. (2015). Prosody and information status in Italian and German L2 intonation. In Chini, M. (a cura di), *Il parlato in italiano L2: aspetti pragmatici e prosodici.* Milano: Franco Angeli. 93-116.

Avesani, C., Vayra, M. (2004). Focus ristretto e focus contrastivo in italiano. In Albano Leoni, F., Cutugno, F., Pettorino, M & Savy, R. (a cura di), *Il parlato italiano, Atti del Convegno nazionale di Napoli,* Napoli, 13-15 febbraio 2003, M. D'Auria Editore, 1-20.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Baumann, S. (2005). Degrees of givenness and their prosodic marking. In *International Symposium on Discourse and Prosody as a complex interface,* Aix-en-Provence.

Baumann, S., Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In Elordieta, G., Prieto, P. (a cura di), *Prosody and Meaning*. Berlin, Boston: De Gruyter Mouton, 119-162.

Belletti, A. (2015). The Focus Map of Clefts: Extraposition and Predication. In Shlonsky, U. (a cura di) *Beyond Functional Sequence: The cartography of Syntactic structures*, v. 10, Oxford: Oxford University Press.

Benazzo, S., Andorno, C. (2017). Is it really easier to acquire a closely-related language? A study on the expression of iteration and continuation in L2 French. In Howard, M., Leclercq, P. (a cura di), *Tense-aspect-modality in a second language: Contemporary perspectives*, 105–143. Amsterdam – Philadelphia: Benjamins.10.1075/sibil.50.05ben

Benazzo, S., Dimroth, C., Santiago, F. (2021). Additive linking in L2 French discourse by German learners: syntactic embedding and intonation patterns. In *Languages,* 6(1).

Bertinetto, P.M. (1981). *Strutture prosodiche dell'Italiano: accento, quantità, sillaba, giuntura, fondamenti metrici.* Firenze: Accademia della Crusca.

Best, C.T. (1995). Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research. In Strange, W. (a cura di), Speech perception and linguistic experience: Issue in cross-language research. Timonium, MD: York Press, 171-204.

Bocci, G., Avesani, C. (2011). Phrasal prominences do not need pitch movements: post-focal phrasal heads in Italian. In Cosi, P., De Mori, R., Di Fabbrizio, G. & Pieraccini, R., (a cura di), *Proceedings of Interspeech* 2011, Firenze, 27-31 August 2011, International Speech Communication Association, 1357-1360.

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [Computer program] Version 6.1.40. https://www.praat.org

Boula de Mareüil, P., Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. In *Phonetica*, 63(4), 247-67.

Brunetti, L., Bott, S., Costa, J., and Vallduví, E. (2011). A multilingual annotated corpus for the study of information structure. In Konopka, M., Kubczak, J., Mair, C., Sticha, F., and Wassner, U. (a cura di), Proceedings of Grammar and Corpora 2009, Third international Conference, Mannheim, Allemagne. Gunter Narr Verlag, Tübingen.

Büring, D. (2016). *Intonation and meaning*. Oxford: Oxford University Press.

Cole, J., Hualde, J.I., Smith, C.L., Eager, C., Mahrt, T., de Souza, R.N. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. In *Journal of Phonetics,* 75, 113-147.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* New York: Cambridge University Press.

Cruschina, S. (2021). The Greater the Contrast, the Greater the Potential: On the Effects of Focus in Syntax. In *Glossa: A Journal of General Linguistics*, 6(1).

Cruttenden, A. (2006). The de-accenting of given information: A cognitive universal? In Bernini, G., Schwartz, M.L. (a cura di), *Pragmatic Organization of Discourse in the Languages of Europe*. Berlin, Boston: De Gruyter Mouton, 311-356.

De Cesare, A.M., Garassino, D. (2015). On the status of exhaustiveness in cleft sentences: An empirical and cross-linguistic study of English also-/only-clefts and Italian anche-/solo-clefts, In *Folia Linguistica*, vol. 49, no. 1, 1-56.

De Cesare, A.M., (2017). Cleft Constructions in romance: definition, types, and parameters of variation. In Dufter, A., Stark, E. (a cura di), Manual of Romance Morphosyntax and Syntax, Berlin; Boston: De Gruyter, pp. 536-568.

De Paolis, B.M., Abbà, B., Romano, A. (2023). On the role of glottal stop: from boundary marker to correlate of focus. A study on Italian and French. In *ICPhS 2023 - 20th Congress of Phonetic Sciences*, Prague.

De Paolis, B.M., Santiago, F., Andorno, C. (2022) Syntaxe ou prosodie? Une étude préliminaire sur l'expression de la focalisation étroite par les apprenants italophones de français L2. *Proc. XXXIVe Journées d'Études sur la Parole - JEP 2022,* Nantes, France.

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., Yoo, H.Y. (2015). Intonational phonology of French: Developing a ToBI system for French. In Frota, S., Prieto, P. (a cura di), *Intonation in Romance.* Oxford: Oxford: Oxford University Press.

Delattre, P. (1966). Les dix intonations de base du français. *The French Review*, 40(1):1–14.

Destruel, E., DeVeaugh-Geiss, J. (2018). On the interpretation and processing of exhaustivity: Evidence of variation in English and French clefts. *Journal of Pragmatics*, 138, 1-16.

Di Cristo, A. (2016). *Les musiques du français parlé: Essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain*, Berlin, Boston: De Gruyter.

Di Cristo, A., Hirst, D. (1996). Vers une typologie des unités intonatives du français. Actes des *XXIème Journées d'étude sur la Parole,* Avignon, France, 219–222.

Dufter, A. (2009). Clefting and Discourse Organization: Comparing Germanic and Romance. In Dufter, A., Jacob, D. (a cura di), *Focus and Background in Romance Languages*, Amsterdam: John Benjamins Publishing Company, pp. 83-121.

Eckman, F. (1981). On the naturalness of interlanguage phonological rules. In Language learning, 31, 1, 195-216. https://doi.org/10.1111/j.1467-1770.1981.tb01379.x

Feldhausen I., Del Mar Vanrell M. (2016). Prosody, Focus and Word Order in Catalan and Spanish: An Optimality Theoretic Approach, *10th International Seminar on Speech Production.*

Féry, C. (2001). Focus and phrasing in French. In Féry, C., Sternefeld, W. (a cura di) *Audiatur vox sapientiae: A festschrift for Arnim von Stechow*, Berlin: Akademie-Verlag. 153–181.

Féry, C. (2013). Focus as prosodic alignment. In *Natural Language & Linguistic Theory*, 31, 3, 683-734.

Flege, J.E. (1995). Second-language speech learning: Theory, findings, and problems. In Strange, W. (a cura di), *Speech perception and linguistic experience: Issue in cross-language research*. Timonium, MD: York Press, 229–273.

Gabriel, C. (2010). On Focus, Prosody, and Word Order in Argentinian Spanish. A Minimalist OT Account. In *Revista Virtual de Estudos da Linguagem*, Special issue 4, 183-222.

Gili Fivela, B. (2012). The perception of intonation in L1 and L2: on phonetic variation and function interpretation. *Proceedings of ML2P: Methods in L2 Prosody, 2012*, *Università di Padova.*

Gili Fivela, B., & Avesani, C. (2021). Analysing Prosody: Methods, issues, and hints on crosslinguistic comparison and L2 learning. In Bernini, G., Valentini, A., Saturno, J., Spreafico, L. (a cura di), *Superare l'evanescenza del parlato. Un vademecum per il trattamento digitale di dati linguistici.* Bergamo: Bergamo University Press, 71–126.

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., Giordano, R., Marotta, G., Savino, M., Sorianello P. (2015). Varieties of Italian and their intonational phonology. In Frota, S., Prieto, P. (a cura di), *Intonation in Romance*. Oxford: Oxford University Press.

Givón, T. (1989). The pragmatics of word order: predictability, importance and attention. Studies in Syntactic Typology (Typological Studies in Language 17). Eds. M. Hammond. Amsterdam and Philadelphia: John Benjamins. 243-284.

Greif, M., Skopeteas, S. (2021). Correction by Focus: Cleft Constructions and the Cross-Linguistic Variation in Phonological Form. In *Frontiers in Psychology,* 12.

Halliday, M.A.K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Hertel, T.J. (2003). Lexical and discourse factors in the second language acquisition of Spanish word order. In *Second Language Research* 19(4), 273-304.

Jarvis, S., Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition.* New York/London: Routledge.

Jun, S.A. (a cura di) (2005). *Prosodic typology: the phonology of intonation and phrasing.* Oxford: Oxford University Press.

Jun, S.A., Fougeron, C. (1995). The Accentual Phrase and the Prosodic structure of French. *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden, vol. 2, 722-725.

Jun, S.A., Fougeron, C. (2000). A Phonological Model of French Intonation. In Botinis, A. (a cura di) *Intonation. Text, Speech and Language Technology,* vol 15. Springer, Dordrecht.

Kisler, T., Reichel, U.D., Schiel, F. (2017): Multilingual processing of speech via web services. In *Computer Speech & Language*, 45, 326–347.

Kiss, K.E. (1998). Identificational focus and information focus. In *Language*, 74, 245–273.

Klein, W., Dimroth, C. (2009). Untutored second language acquisition. In Ritchie, W.C., Bhatia, T.K. (a cura di) *The new handbook of second language acquisition*. Bingley: Emerald, 503-522.

Klein, W., Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). In *Second Language Research*, 13(4), 301–347.

Krifka, M. (2008). Basic notions of information structure. In *Acta Linguistica Hungarica*, 55(3–4), 243–276.

Lambrecht, K. (1994). *Information structure and sentence form: Topics, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

Lambrecht, K. (2001). A Framework for the Analysis of Cleft Constructions. In *Linguistics*, 39, 463-516.

Lambrecht, K. (2001). A framework for the analysis of cleft constructions. In *Linguistics* 39(3). 463–516.

Larrivée, P. (2022). The curious case of the rare focus movement in French. In Garassino, D., & Jacob, D., (a cura di), *When Data Challenges Theory: Unexpected and paradoxical evidence in information structure,* Amsterdam: John Benjamins, 183-202.

Lozano, C. (2006). Focus and split-intransitivity: the acquisition of word order alternations in non-native Spanish. In *Second Language Research*, 22(2), 145-187.

Mack, M. (1988). Interlanguage Phonology: The Acquisition of A Second Language Sound System. In Ioup, G., Weinberger, S.H. (a cura di), *Studies in Second Language Acquisition*, 10(3), 426-428.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. In *Journal of Phonetics*, vol. 32, 543-563.

Mennen, I. (2015). Beyond Segments: Towards a L2 Intonation Learning Theory. In Delais-Roussarie, E., Avanzi, M., Herment, S. (a cura di.), *Prosody and Language in Contact. Prosody, Phonology and Phonetics.* Springer, Berlin, Heidelberg.

Mennen, I., & de Leeuw, E. (2014). Beyond Segments: Prosody in SLA. In *Studies in Second Language Acquisition*, 36(2), 183–194.

Mertens, P. (2012). La prosodie des clivées. Caddeso, S., Roubaud, M.N., Rouquier, M., Sabio, F., (a cura di), *Penser les langues avec Claire Blanche Benveniste*. Aix-en-Provence: Presses Universitaires de Provence.

Mertens, P. (2012). La prosodie des clivées. In Caddeo, S., Roubaud, M.N., Rouquier, M., Sabio, F. (a cura di.) *Penser les langues avec Claire Blanche Benveniste*. Aix-en-Provence: Presses Universitaires de Provence, 127-139.

Mertens, P. (2014). Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. In *Journal of Speech Sciences,* 4, 2, 17-57.

Michelas, A., German, J. (2020). Focus marking and prosodic boundary strength in French. In *Phonetica*, 77, 244-267.

Neeleman, A., van de Koot, H. (2014). Word Order and Information Structure. In Féry, C., Ishihara, S. (a cura di), *The Oxford Handbook of Information Structure*, Oxford: Oxford University Press.

Pinelli, M.C., Poletto, C., Avesani, C. (2020). Does prosody meet syntax? A case study on standard Italian cleft sentences and left peripheral focus. *The Linguistic Review*, 37, no. 2, 309-330.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Repp, S. (2016). Contrast : Dissecting an Elusive Information-Structural Notion and Its Role in Grammar. In Féry, C., Ishihara, S. (a cura di), *The Oxford Handbook of Information Structure*, Oxford: Oxford University Press, 1–23.

Romano, A., Interlandi, G. (2008). La variazione geoprosodica nelle varietà di italiano parlate a Torino, Milano, Roma e Palermo: come l'intonazione veicola l'informazione pragmatica. In: Turculet, A. (a cura di) *La variation diatopique de l'intonation dans le domaine roumain et roman*. Iasi: Editura Universitatii Al. I. Cuza, 185-215.

Rooth, M.A. (1992). A theory of focus interpretation. In *Natural Language Semantics*, 1, 75–116.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. In *Phonetica*, 23, 1–33.

Samo, G., Merlo, P. (2021). Intervention effects in clefts: a study in quantitative computational syntax. *Glossa: a journal of general linguistics*, 6(1):145.

Sbranna, S., Ventura, C., Albert, A., Grice, M. (2023). Prosodic marking of information status in Italian. In *Journal of Phonetics,* 97. Preprint.

Selinker, L. (1972). Interlanguage. In *International Review of Applied Linguistics in Language Teaching,* 10, 209-241.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: The MIT Press.

So, C.K.J., Best, C.T. (2008). Do English speakers assimilate Mandarin tones to English prosodic categories? *Proceedings of Interspeech 2008*, 1120.

Turco, G., Braun, B., Dimroth., C. (2015). Prosodic and Lexical Marking of Contrast in L2 Italian. In *Second Language Research*, 31(4): 465–91.

Vallduví, E., Engdahl, E. (1996). The linguistic realisation of information packaging. In *Linguistics*, vol. 34, no. 3, 459-520.

Vander Klok, J., Goad, H., Wagner, M. (2018). Prosodic focus in English vs. French: A scope account. In *Glossa: a journal of general linguistics*, 3(1): 71.

Vedder, I. (2008). Competenza pragmatica e complessità sintattica in italiano L2: l'uso dei modificatori nelle richieste. In *Linguistica e Filologia*, 25, 1.

Wiener, S., Goss, S. (2018). Perceptual assimilation of non-native prosodic cues: Cross-linguistic effects of lexical F0 learning. *Proceedings of Speech Prosody 2018*, 947-951.

Yan, M., Warren, P., Calhoun, S. (2022). Focus interpretation in L1 and L2: The role of prosodic prominence and clefting. *Applied Psycholinguistics*, 43(6): 1275-1303.

Zubizarreta, M.L., Nava, E (2011). Encoding discourse-based meaning: Prosody vs. syntax. Implications for second language acquisition. *Lingua*, 121, 652-669.

PARTE IV

RICONOSCIMENTO AUTOMATICO DEL PARLATO E DEL PARLANTE

AUTOMATIC SPEECH AND SPEAKER RECOGNITION

DOI: 10.17469/O2111AISV000024

ELEONORA RODEGHER, ALESSIO BRUTTI, DANIELE FALAVIGNA

Automatic speech recognition for low-resource languages: exploring transfer learning from related languages

Automatic speech recognition systems typically require huge amounts of data, which are not always available for languages lacking digital resources. In this work we address the problem of automatic speech recognition for low-resourced languages by apply transfer learning approaches to the large-scale pre-trained model wav2vec2-xls-r. In particular, we present a sequential adaptation strategy where, using a common tokenizer, the large-scale model is firstly fine-tuned on a genealogically related well-resourced language ("pivot") and then further adapted on the target low-resourced one. Experimental results in terms of word error rate, show that this approach considerably outperforms models directly fine-tuned on the target language, in particular when very few data are available. Experiments involve Italian, Spanish, Portuguese and German as well-resourced languages and two minority languages: Galician and the variety of Romansh called Vallader.

Keywords: Automatic Speech Recognition, Transfer Learning, Large-Scale pre-trained model, low resourced languages.

1. *Introduction*

Automatic speech recognition (ASR) systems typically require huge amounts of data, which are not always available for languages lacking digital resources. Recently, new approaches have been investigated to overcome the scarcity of labelled data (Besancier, Barnard, Karpov & Schultz, 2014). One method is to adapt models pre-trained on vast quantities of data from well-resourced languages to poorly resourced ones. Differently from what commonly done in literature, i.e. training multilingual models or fine-tuning a large-scale pre-trained model on the target language, in this work we present a transfer learning strategy which fine-tunes the multilingual pre-trained model wav2vec2-xls-r (Babu, Wang, Tjandra, Lakhotia, Xu, Goyal, Singh, Platen, Saraf, Pino, Baevski, Conneau, & Auli, 2021) on different low-resourced languages using a pivot well-resourced one. In particular, we considered four well-resourced languages: Italian, Portuguese, Spanish and German; and two minority languages: Galician and Romansh Vallader.

2. *State of the art on ASR systems for low resource languages*

Recent works on ASR systems for low-resource languages are illustrated in this section. We refer the reader to (Meng, Yolwas, 2022) for a more complete survey. The work in (Docío-Fernandez & Garcia-Mateo, 2002) was one of the first investigations on the use of multiple languages to develop ASR systems for low-resourced languages. The paper proposes a bilingual ASR system for Spanish and Galician, having a scarce amount of data for Galician, training a multilingual model on a global set of phonemes. A more recent study (Yi, Wang, Cheng, Zhou, & Xu, 2021), used limited speech data from real spoken scenarios in six different languages to adapt the wav2vec2.0 model pretrained on English (Baevski, Zhou, Mohamed, & Auli, 2020) The authors evaluated different speech units and showed that using sub-words or characters led to slightly better results. More analogous to our approach, (Bermuth, Poeppel, & Reif, 2021) demonstrated the benefit of transfer learning where language similarities occur. A QuartzNet architecture, pre-trained for English is adapted to German and other languages with a similar alphabet. Similarly, in (Jacobs, Kamper, 2021) transfer learning on genealogically related languages is investigated by training multilingual word embedding models on South African low-resourced languages. The results proved that training a multilingual model using closely related languages improves the model's performance. Finally, a transliteration transfer learning approach was tested in (Khare, Mittal, Diwan, Sarawagi, Jyothi, & Bharadwaj, 2021), converting the transcriptions of a high-resource language using the graphemes of low-resource ones. The languages involved were English and six low-resourced, e.g. Hindi, Telugu, Gujarati, Bengali, Korean, and Amharic. The model wav2vec2 was pre-trained on these transliterations and fine-tuned on limited speech data of the other languages.

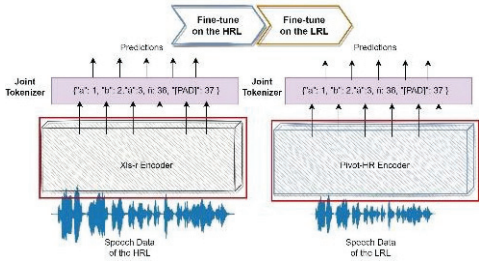
Although our approach relies on a pre-trained large-scale model as wav2vec2.0-xls-r and on similarities between languages, as many of the works mentioned above, we propose a novel two step strategy which uses the well-resource language for a first adaptation of the pre-trained model.

3. *Proposed Approach*

A common approach to facing the sparseness of data is to use transfer learning. This technique consists in reusing a model trained on one task for another similar or related task. In ASR domains, transfer learning techniques can be employed for transferring the knowledge from a model trained on speech from high- resource languages to low-resourced ones, by fine-tuning all or parts of the model on speech from the low-resourced ones (Kahre *et al.*, 2021). This permits to reuse the speech representations shared among languages when fine-tuning on the target language. The pre-trained model wav2vec2-xls-r is the large-scale multi-lingual version of wav2vec2.0. The model was pre-trained on half a million hours of speech from 128 languages, using the self-supervised learning approach. For ASR tasks, a linear layer is added on top of the pre-trained model to predict the output tokens using the connectionist temporal classification (CTC) loss (Graves, Fernandez, Gomez, & Schmidhuber, 2006).

Being able to provide representations for multilingual speech inputs wav2vec2-xls-r is a promising candidate for applying transfer learning for a specific low-resourced language. Nevertheless, when the amount of training data is extremely low, a direct fine tuning of the pre-trained multi-lingual model is not sufficiently effective. In this work we propose a 2-step approach: first the pre-trained model is fine-tuned on a language genealogically similar for which enough training material is available (we refer to this as the pivot language); then the fine-tuned model is further adapted on the data of the low-resourced target language. Figure 1 shows a graphical example of the proposed approach applied to Galician with Italian as pivot language.

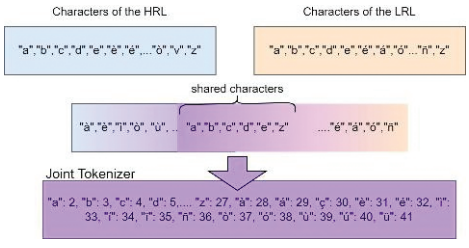
Figure 1 - *Example of the proposed approach applied to Galician with Italian as pivot language. The wav2vec2.0-xls-r model is firstly adapted on the Italian data. Then, the fined tuned model is further trained using the Galician data*



Clearly, the choice of the pivot language is fundamental as it should share acoustic and graphemic similarities with the target language.

A relevant aspect in the implementation of the proposed approach is the definition of the tokenizer to be used on top of the wav2vec2.0 encoder. Given that we use graphemes as tokens of our speech recognition system, we utilize a joint tokenizer obtained by merging the graphemes of the pivot and target languages. This implementation choice allows preserving the knowledge related to the common graphemes and use the small training set of the low-resourced language to learn the specific graphemes. Figure 2 shows an example of joint tokenizer.

Figure 2 - *Example of the creation of the joint tokenizer*



4. *Dataset*

The dataset adopted in our analysis was Version 8 of the Mozilla Common Voice by (Ardila, Branson, Davis, Kohler, Meyer, Henretty, Morais, Saunders, Tyers, & Weber, 2020). The open crowd-sourced multi-language dataset, as of August 2022, counts 96 languages. Users can contribute to enriching the corpus of a specific language by recording their voice or evaluating the transcriptions of already recorded speech data. The audio files are in MPEG-3 format with a 48kHz sampling rate (Ardila *et al.*, 2020). The recordings are primarily read sentences from Wikipedia. The data are complemented with a rich set of information regarding the amount of validated audio files, the overall duration, and the average length of an audio file. Each corpus is split into the train, test, and validation sets. The TSV file contains information about the speakers, such as their age, sex, and accent, the path to the audio files, and the transcriptions of the read sentences. Speech data and transcriptions have been evaluated and only clips marked as valid are included in the official training, development, and testing sets for each language.

In our study, for each corpus the train set was used to fine-tune the models, while we made use of the validation set to test the model in training. The test set was passed to the fine-tuned models for inferences, as these data were never passed to the model in training. As mentioned above, we extracted 4 well-resourced languages: Italian, Spanish, Portuguese and German and 2 very low-resourced languages: Galician and Romansh Vallader. More details about the data of the latter two languages are reported later.

5. *Experimental analysis*

The experiments considered different amounts of data to better assess on the impact of the pivot language. Performance was evaluated in terms of Character and Word Error Rates.

The code is available here: https://github.com/elerdg/ASR-for-low-resource-languages. We point the reader to that repository for details about the hyperparameters and the data pre-processing. As baselines, we created monolingual models for Galician, and Romansh Vallader, considering for each language a tokenizer containing all the characters plus special tokens for CTC decoding. Note that specific training strategies, as for example data augmentation, can be used to improve the performance and the generalization capabilities of the model. Since we are interested in investigating the impact of using a pivot language, we do not apply any of these methods.

5.1 Experiments on Galician

Our first experiment involves the Galician, a minority language spoken in the Spanish region of Galicia situated in the northwest of the Iberic peninsula. This language is part of the romance language family and has historical roots with Portuguese (Ramallo, 2015).

Version 8 of Mozilla Common Voice contains 10.17 validated hours of Galician, equivalent to 7618 sentences, from 486 different speakers. The average duration of an audio file is 4.808 seconds. In the train set there are 3062 files which are equal to 4 hours of speech, in the test and the validation set there are 2258 and 2240 respectively, both correspond to approximately 3 hours of total duration.

5.1.1 Tokenizer

The vocabulary of the language contains: 26 characters of the alphabet, 7 letters marked by diacritics “á”, “é”, “í”, “ó”, “ú”, “ï”, “ü” and the apostrophe. Moreover, not knowing which standard was adopted in the transcriptions of the read sentences, we included the grapheme “ç”.

5.1.2 Results

Figure 3 - *CER on the Galician test set when fine-tuning using the Galician data only (orange line) and when using Italian, Spanish, Portuguese and German as pivot, considering different amount of data for fine- tuning*

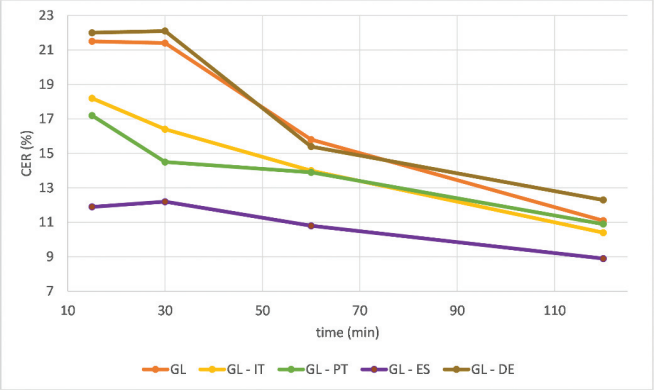
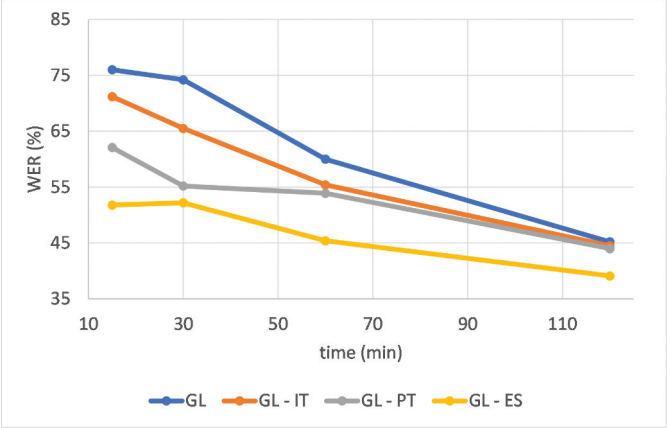


Figure 3 reports the CER obtained on Galician considering different amounts of material for training and different pivot languages. The orange line is our baseline consisting of the pure adaptation of wav2vec2.0-xls-r on the Galician data. Note that using Spanish (purple line), as expected, leads to the best performance, since it shares a large number of graphemes and have high overlap of phonemes, with more than 10 percent point reduction with respect to the baseline. Italian and Portuguese also give a clear improvement but rather inferior than Spanish. Finally, German, being a language from a different family with different phonology and orthography, brought no advantage at all, providing very similar performance as the monolingual model obtained using only Galician data. Note that the number of tokens is equivalent in all the experiments. Moreover, the performance of the Spanish system on Galician is above the baseline (around 30 % CER). Clearly, as the amount of training data increases, the gap between the different systems is reduced as the model tends to forget what learned in the first adaptation step. Note,

however, that Spanish is remarkably better than the other languages. Fig. 4 reports the results using WER as metric. Note that we are not using any language model.

Figure 4 - *Performance in terms of WER for Galician. No language model is used*



5.2 Experiments on Vallader

Spoken in the southeastern part of Switzerland, Romansh is a romance language belonging to the Rhaeto-Romance branch. In 1938 it was recognized as one of the four national languages of the country, together with German, Italian, and French. The origin and its classification inside the romance family is a debated theme among linguists.

The dataset available for the Romansh Vallader variety in Common Voice is extremely small since it contains only 2.34 hours of labelled validated data. The total number of the audio files is 1459 divided into the train, test, and validation sets. The average duration per audio file is 5.791 seconds. The number of speakers in this corpus is 50.

5.2.1 Tokenizer

To construct the tokenizer, we considered the convention adopted to transcribe the audio files, which in our case is the writing system of the Vallader variety. Differently from the standard Rumantsch Grischun, Vallader transcribes three vowels with the dieresis: “ï”, “ö”, “ü”. The vocabulary of the tokenizer contains in total 39 characters, of which five vowels accented with the grave accent “à”, “è”, “ì”, “ò”, “ù”, three vowels with the circumflex accent “â”, “ê”, “ô”, and three vowels with dieresis “ï”, “ö”, “ü”. We inserted “é” due to the possible presence of loan words from Italian that may present an acute accent.

5.2.2 Results

Figure 5 - *Performance in terms of CER on Romansh Vallader using Italian as pivot*

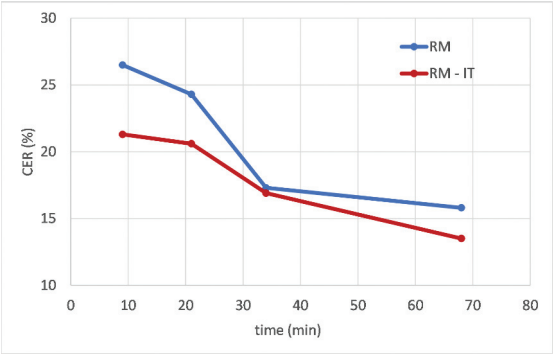


Fig. 5 shows the CER performance on Romansh Vallader with and without pivoting on the Italian language. Similar trends can be observed as for the Galician language confirming the benefit of using a well-resourced language for a first fine-tuning of the pre-trained model. Remarkably, note that the performance is similar to those of Galician using Italian when similar amounts of training data are considered.

6. *Conclusions*

In this study we presented a transfer learning approach to obtain an ASR system for very low-resourced languages using a well-resourced language as pivot. The type of language selected as the pivot was found to be influential. In the case of Galician, we observed the effect of using different romance languages. The better performances were obtained using the languages belonging to the same branch and in contact to Galician. Spanish was proven to be the best choice: the languages share a large number of graphemes and have high overlap of phonemes. The phonological differences between the two languages were demonstrated by testing the Spanish model on Galician which scored 15 % CER. Fine-tuning this model on 2 hours of the low-resource language obtains 8.9 % CER, proving the advantages of transfer learning. The experiment using German confirmed the role of the sharing between the two languages. Experiments on Romansh Vallader and Italian confirmed the trends observed for Galician

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A Massively Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference 2020*, Marseille, France, 4218-4222.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Platen, P.V., Saraf, Y., Pino, J.M., Baevski, A., Conneau, A., & Auli, M. (2022) XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of 23rd Interspeech Conference*, Incheon, Korea, 18-22 September 2022, 2278-2282.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems* (NeurIPS 2020), Vancouver, Canada, 12449-12460.

Bermuth, D., Poeppel, A., & Reif, W. (2021). Scribosermo: Fast Speech-to-Text models for German and other Languages. ArXiv 2021, ArXiv:2110.07982.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. In *Speech Communication*, 56, 85–100.

Docio-Fernandez, L., & Garcia-Mateo, C. (2002). Acoustic Modelling and Training of a Bilingual ASR System when a Minority Language is Involved. In *Proceedings of the International Conference on Language Resources and Evaluation*, Gran Canaria, Spain, May 2002, 3, 873-876.

Graves, A., Fernandez, S., Gomez, F. & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning 2006* (*ICML)*, Pittsburgh, Pennsylvania, USA, June 25-29, 369–376.

Jacobs, C., & Kamper, H. (2021). Multilingual Transfer of Acoustic Word Embeddings Improves When Training on Languages Related to the Target Zero- Resource Language. In *Proceedings of the 22nd Interspeech Conference 2021*, Brno, Czech Republic, 30 August – 3 September, 1549-1553.

Khare, S., Mittal, A., Diwan, A., Sarawagi, S., Jyothi, P., & Bharadwaj, S. (2021). Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proceedings of the 22nd Interspeech Conference 2021*, Brno, Czech Republic, 30 August – 3 September, 1529-1533.

Meng, W. & Yolwas, N. (2022). A Review of Speech Recognition in Low-resource Languages. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Learning*, Chengdu, China, 245-252

Ramallo, F. (2007). Sociolinguistics of Spanish in Galicia. In *International Journal of the Sociology of Language*, 184, 21-36.

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages. ArXiv:2012.12121.

DOI: 10.17469/O2111AISV000025

FRANCESCO SIGONA, MIRKO GRIMALDI

Calibration and Fusion of Parametric and Non-parametric Statistical Methods for Forensic Semi-Automatic Speaker Recognition Systems: experimental results

In this paper, the most employed Forensic Semi-Automatic Speaker Recognition (FSASR) approaches used in Italy – as emulated in the IMPAVIDO software, and some variants of them – are investigated from the point of view of the accuracy. In doing this, we employed score calibration and fusion with Logistic Regression (LogReg). The metric used to assess the accuracy was the likelihood-ratio cost (Cllr), while the factors of analysis include the database numerosity, different software implementation of LogReg, and the strategy to consider the features available for the different vowels. Our results show that the LogReg calibration has systematically improved the Cllr, suggesting that also the accuracy of the tested FSASR approaches may be improved introducing calibration. Furthermore, our findings suggest that the tested FSASR approaches give more accurate results when the IMPAVIDO database is used.

Keywords: Accuracy, Likelihood-Ratio, Bayesian framework, Forensic Voice Comparison.

1. *Introduction*

In Forensic Speaker Recognition (FSR), the recording of an anonymous speaker is compared with a recording of a speaker whose identity is known. The aim is to help a court of law to decide if the two recorded voices could belong to the same speaker (Morrison, Enzinger, Phil and Zhang, 2017). Many approaches have been proposed to achieve FSR: nonparametric-subjective approaches such as auditory, spectrographic (also known as "voicegram" identification, or "voiceprinting": Kersta, 1962; Tosi, Oyer, Lashbrook, Pedrey, Nicol and Nash, 1972), and parametric-objective approaches, which are based on quantitative measurements and analysis of specific features of the speech signal (Rose, 2002; Rose, 2006) particularly those associated with evidence, are exemplified and critically discussed, and comparisons drawn with generic Speaker Recognition. The centrality of the Likelihood Ratio of Bayes' theorem in correctly evaluating strength of forensic speech evidence is emphasised, as well as the many problems involved in its accurate estimation. It is pointed out that many different types of evidence are of use, both experimentally and forensically, in discriminating same-speaker from different-speaker speech samples, and some examples are given from real forensic case-work to illustrate the Likelihood Ratio-based approach. The extent to which Technical Forensic Speaker Recognition meets the Daubert requirement of testability is

also discussed. © 2005 Elsevier Ltd. All rights reserved.","author":[{"dropping-particle":"","family":"Rose","given":"Phil","non-dropping-particle":"","parse-names":false,"suffix":""}],"container-title":"Computer Speech and Language","id":"ITEM-1","issue":"2-3 SPEC. ISS.","issued":{"date-parts":[["2006"]]},"page":"159-191","title":"Technical forensic speaker recognition: Evaluation, types and testing of evidence","type":"article-journal","volume":"20"},"uris":["http://www.mendeley.com/documents/?uuid=53443a14-a72e-4773-9563-56b19157e8a9"]}],"mendeley":{"formattedCitation":"(Rose, 2006; Jessen, 2008). The objective approaches, in turn, can be divided in acoustic-phonetic, or semi-automatic approaches, and automatic approaches. In the former, the expert plays a key role in the choice of the phonetic units and the extraction of the features to be compared. In the automatic approaches, these steps are mostly accomplished by the software program that also perform the voice comparison.

About the output of automatic and semi-automatic approaches, in recent years the Likelihood Ratio (LR) framework has been proposed as a logical way to report evidence to the court of law (Aitken, Taroni, 2004): furthermore, it has been recommended by the European Network of Forensic Science Institutes (ENFSI)[1] as the ideal method to use in FSR. The LR represents a balance between the probability to get the evidence given the prosecution hypothesis (or same-origin hypothesis, e.g., the questioned material has been generated by the suspect) and the probability to get the same evidence given the defence hypothesis (or different-origin hypothesis, e.g. the questioned material has been generated by someone else). Therefore, two important properties can be recognized to the LR:

1. The LR represents the degree of similarity of a pair of samples (the two recordings), while taking into consideration their typicality with respect to a model of the relevant population, so that a more positive value indicates stronger support for the voices being from the same individual (and vice-versa)
2. The LR value has a meaning by itself, in the sense that a LR value greater than 1 supports the prosecution hypothesis and, more precisely, a LR value equals to 3 (for instance) means that the evidence supports the prosecution hypothesis with a degree of 3 versus 1 (Ramos, Krish, Fierrez and Meuwly, 2017).

In general, the output of a FSR system is (at least) a number, namely a "score", that may have the first property of LR (in this sense, a score may be LR-like) but does not necessarily also have the second property, i.e. it does not have a clear probabilistic interpretation. Indeed, these scores may contain some bias due to condition mismatch between the data that the system or speaker model was trained on and the conditions of the verification speech. Therefore, after calculating the scores, a further step may be necessary to convert them into true LRs, to be compliant with the LR framework: this step is called *calibration*. It is important to note that the same term, *calibration*, is often used also to indicate a property of a set of LR values, which can be measured, and it is strictly related to the accuracy of a FSR system (Ramos, Gonzalez-Rodriguez, 2013) an incorrect selection of databases, a bad

[1] http://www.enfsi.eu.

choice of statistical models, low quantity and bad quality of the evidence are factors that may lead to likelihood ratios supporting the wrong proposition in a given case. However, measuring performance of LR values is not straightforward, and adequate metrics should be defined and used. With this objective, in this work we describe the concept of calibration, a property of a set of LR values. We highlight that some desirable behavior of LR values happens if they are well calibrated. Moreover, we propose a tool for representing performance, the Empirical Cross-Entropy (ECE.

If the same casework can be processed by two or more FSR systems, their output scores may be combined into a single outcome. This can be achieved by a *fusion* procedure, which aims to produce a compound system that performs better than the single ones.

Score calibration methods based on generative models have recently gained interest (van Leeuwen, Brümmer, 2013; Brümmer, Swart and van Leeuwen, 2014; Plchot, Matějka, Silnova, Novotny, Diez, Rohdin and Glembek, 2017; Cumani, Laface, 2019), but the most widely adopted method is still the score calibration by means of Logistic Regression (Morrison, 2013), a well know and largely applied technique in machine learning, developed for automatic speaker recognition and then become popular in FSR.

The aim of this study is to evaluate the performance of two FSASR approaches traditionally used in Italy, adopting LogReg techniques applied to calibration and fusion. The two systems referred to are IDEM/SPREAD (first appeared in Falcone *et al.*, 1992; see Sigona and Grimaldi, 2015 for a discussion and further references), and SMART (Bove, Giua, Forte and Rossi, 2002; Jona Lasinio, Rossi and Bove, 2004; Romito, Bove, Delfino, Rossi, and Jona Lasinio, 2009). The main difference between the two systems is that IDEM/SPREAD is based on parametric (multivariate Gaussian) statistical models, while SMART is based on non-parametric statistical models.

Recently, both approaches have been integrated into eIDEM and eSMART emulators that are the two main components of a new tool for FSASR, named IMPAVIDO (Sigona, Grimaldi, 2017), and which is used for this study, together with some related variants under development.

The remaining of this paper is organized as follows. Section 2 recalls some fundamentals about LogReg in terms of score calibration and fusion. Section 3 describes the score computation by eIDEM, eSMART and their variants employed for the current study. Sections 4 and 5 describe the experimental setup and present the achieved results. Finally, in Section 6 the main findings are discussed, and future research directions are highlighted.

2. *Logistic Regression and score calibration*

As shown in formulas (1) to (3), if p is the probability of the prosecutor hypothesis, then the LR can be expressed as odds of p (1), the logarithm of LR (let LLR) is the log-odds of p (2), and p is related to LLR by means of the logistic function (3):

$$(1) \quad LR = \frac{p}{1-p}$$

$$(2) \quad LLR = \log LR = \log \frac{p}{1-p}$$

$$(3) \quad p = \frac{1}{1+e^{LLR}}$$

If p>0.5, then necessarily 1-p<0.5, then we say that there is more support for the prosecution hypothesis than the defence one. In this case, is straightforward to observe that LR must be greater than 1, and this is why true LRs have a clear probabilistic interpretation, while raw uncalibrated "scores" do not (a score merely greater than 1 does not necessarily means that p>0.5, i.e. it does not necessarily support the prosecution hypothesis).

What we need now is a function that maps the LR-like scores output by FSR systems to LRs: this is where LogReg comes in. According to LogReg, the following affine relation in the log-odds space gives the score-to-LR transformation:

$$(4) \quad \log(LR) = w_0 + w_1 s$$

where s is the logarithm of the score (log-score), w_0 is a shift coefficient and w_1 is a scale coefficient that must be estimated by means of an appropriate training.

LogReg can also be applied to calibration and fusion of multiples scores coming from different FSR systems, which may process, with their own specificities, the *same* couple of voice segments. In this case, the transformation in (4) must be slightly changed by keeping the shift factor w_0 and including as many w_i scale coefficients as the number of scores (i.e. FSR systems) to fuse.

In both cases of calibration and calibration plus fusion, the required training data consist of a set of scores from comparisons where some are known to be *same-origin comparisons* and others are known to be *different-origin comparisons*, taking the speakers data from the database of the relevant population (at least two different sessions for each speaker are required to make same-speaker comparisons). The estimation of the LogReg coefficients is usually done by means of iterative procedures that minimize a cost function. In the current study, prior probabilities are implicitly set to 0.5 (prior odds = 1, i.e. there is no a-priori preference for the prosecution hypothesis with respect to the defence hypothesis). Hence, the cost function to be minimised, i.e. the deviance statistic, is equivalent to the log-likelihood-ratio cost Cllr given by the following equation (Brümmer, Du Preez, 2006; van Leeuwen, Brümmer, 2007):

$$(5) \quad C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}} \log_2\left(1+\frac{1}{LR_{so_i}}\right) + \frac{1}{N_{do}}\sum_{j=1}^{N_{do}} \log_2(1+LR_{do_j})\right)$$

where N_{SO} and N_{DO} are the number of same-origin and different-origin comparison, $LR_{SO}i$ and $LR_{SO}i$ are the LR values for same-origin and different-origin comparisons.

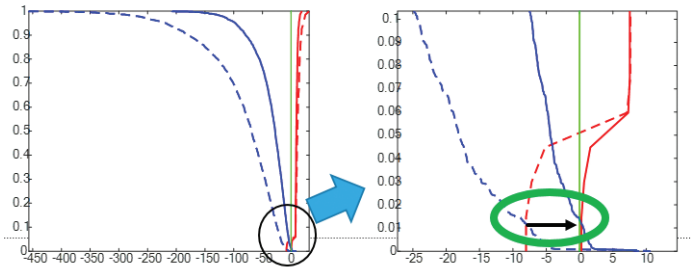
Cllr represents the total probability of error, at any possible prior, that the fact finder would obtain by using the LR value computed by the forensic expert,

in the Bayesian decision framework. Thus, the less the Cllr, the more accurate is the system, and the Cllr computed on calibrated scores, i.e. on LLR, values, is expected to be smaller than the Cllr computed on uncalibrated raw scores. In other words, Cllr can be used as an index to assess the performance of calibration and fusion, i.e. the FSR system, as depicted in Figure *1*.

In doing this, different software implementations of LogReg have been tested:

a) the Focal toolkit[2];
b) the Bosaris toolkit (Brümmer & Villiers, 2013).
c) a variant by Morrison[3] of the LogReg training function of the Focal Toolkit (Morrison, 2011);
d) the *mnrfit* Matlab® function for multinomial regression.

Figure 1 - *Tippet plots illustrate the Probability of false identification (P.f.id., depicted by blue lines on the left side of each plot) and Probability of false rejection (P.f.rej., represented by red lines on the right side of each plot) as functions of log-scores (horizontal axis), for both uncalibrated (dashed lines) and calibrated (solid lines) scores. The plot on the right provides a close-up view of the left plot, clearly demonstrating that, following the calibration process, the point where the two error probabilities (and thus the supports of the two competing hypotheses) are equal approaches log-score = 0, as would be anticipated for a LR value (indeed LR=1 equates to LLR = 0). In this specific instance, the Cllr for uncalibrated scores was 0.226, while the Cllr for the same scores after calibration notably improved to 0.034*



3. *The tested FSASR systems.*

In this study, traditional Italian FSASR approaches have been tested, by means of emulators included in the IMPAVIDO software, eIDEM and eSMART. Both emulators, like the original software, rely on pre-extracted features values, namely, the fundamental and the first three formant frequencies (F0, F1, F2, F3) of the Italian vowels. The IDEM/SPREAD software was designed to handle only 4 Italian vowel /a, ɛ, i, ɔ/, and therefore up to 16 features altogether. The vowel /u/, for example, cannot be input in the IDEM/SPREAD software by design. Due to this limitation, even if the IMPAVIDO software can manage a 7-vowel system

[2] Niko Brummer, "Focal toolkit," available at https://sites.google.com-/site/nikobrummer/focal.
[3] The code of Morrison is available at http://geoff-morrison.net.

/a, ɛ, e, i, ɔ, o, u/ (and can also be easily customized by the authors to manage the formants of any additional vowel sound), in the continuation of this work all the test will be conducted on the 4 vowels listed above.

eIDEM is an emulator of IDEM/SPREAD method. According to this method:

a) the between-speaker variability of the features is modelled by a multivariate Gaussian distribution whose centroid (the vector of means) and variance-covariance matrix are estimated from the database of features that comes with the software (or optionally provided by the user him/herself).
b) the within-speaker variability is modelled by a multivariate Gaussian distribution, whose variance-covariance matrix is estimated from the database, and whose centroid is estimated from the values of the suspect features; this also implies the following:
c) the statistical distribution of the (squared) Mahalanobis distance between the features of the anonymous and the suspect speakers can be modelled by a Chi-squared distribution with as many degrees of freedom as the number of acoustic features used for the test (e.g. 16).

Based on the above assumptions, the software calculates:

1. the value of the (squared) Mahalanobis distance between the feature sets of the anonymous and the suspect (let d^2_{AS}).
2. the probability of false rejection (let $P_{f.rej.}$) as the area of the queue of the above Chi-square distribution starting from the value d^2_{AS}.
3. the value of the probability of false identification (let $P_{f.id.}$), as the integral of the Gaussian probability density function (p.d.f.) that models the between-speakers variability, onto the ellipsoid with squared radius equals to d^2_{AS} and centred to the suspect centroid.
4. a distance threshold (let d^2_{th}) as the smaller between the value of the inverse Chi-square (by default, computed at the assigned value α=0.01) and the value of the squared radius of the ellipsoid which would give a $P_{f.id.}$ equals to a second assigned value (by default, 0.001).
5. the response IDENTIFIED and a LR-like score given by the ratio $(1\text{-}P_{f.rej.})/P_{f.id.}$, if $d^2_{AS} > d^2_{th}$, or the response UNIDENTIFIED and a LR (-like) score given by the ratio $(1\text{-}P_{f.id.})/P_{f.rej,}$ otherwise (for these formulas, see Federico, 2010; but also see Sigona & Grimadi, 2015), about some anomalies in some releases of IDEM/SPREAD).

As far as the calibration and fusion experiments is concerned, two variants of eIDEM were introduced (without any attempt to obtain scores already calibrated at this level):

a) eIDEM_NOTH: it does not care about distance threshold (indeed, the "pure" LR framework does not demand any binary decision depending on thresholds), but simply outputs a score computed as $(1\text{-}P_{f.rej.})/P_{f.id}$, i.e. the same formula that eIDEM uses in case of identification.

b) eIDEM_LK: as eIDEM_NOTH, it ignores the threshold and simply outputs a score given by the ratio between the above mentioned Chi2-squared p.d.f. at the d^2_{AS} point and the between-speaker variability p.d.f. at the point given by the anonymous centroid.

The last FSASR method used in the current study is the one implemented by eSMART, which is based on non-parametric statistics described in (Bove *et al.*, 2002). In this case, the LR (-like) result is given by the ratio of the kernel estimates of the p.d.f. modelling the distance between features of the suspect from himself and the p.d.f. modelling the distance between the suspect ad any other speaker of the database, at a point given by the distance between the suspect and the anonymous speakers.

4. *Experiment setup*

As already mentioned, this study is based on F0, F1, F2 and F3 of the Italian /a, ɛ, i, ɔ/ vowels. Two databases of features representative of the Italian reference population have been used.

The first one is the default database of the IDEM/SPREAD software. This resource essentially consists of a single table containing 26642 rows (records), each of which stores a vector (a quadruple) of values of the parameters F0, F1, F2, F3, together with the indication of the corresponding vowel and an additional text string that indicates (anonymously) the respective speaker. The formants calculation was based on environmental or telephone interceptions. Except for knowing that all the speakers in question, 428, are male, no other information about the socio-linguistic composition of this database is available. Therefore, it is an absolutely unstructured database, which, however, has been used in conjunction with the software for years (and probably still is).

The study would certainly have benefited from a comparison between this database and the SMART database, if only for the sake of fairness. However, it seems that accessing the SMART database is not possible. Nevertheless, it was decided to introduce a second database into the tests, similarly unstructured as the IDEM/SPREAD one. This is the default database distributed with IMPAVIDO, and represents the formants of 310 male voices, in similar recording conditions of the IDEM/SPREAD database.

To assess the capability of an FVC system to recognize two recordings of the same speaker, it is necessary to have a database where each speaker is represented by at least two distinct, non-contemporaneous recordings. However, in both available databases, such a distinction is not present. Therefore, for the purposes of this study, to make same-origin comparison it was necessary to split the single feature matrix of each speaker into two equisized feature matrices, simulating two different sessions for each speaker. In the IMPAVIDO database, each speaker has 20 occurrences for each vowel; thus, 2 sessions of 10 occurrences for each vowel for all speakers were simulated. This number of occurrences is quite adequate to

reproduce a real casework situation. In the IDEM/SPREAD database, only 67 speakers out of 428 have at least 20 occurrences for each vowel. Therefore, for the current study, only those 67 speakers were used from that database.

All the required same-origin and different-origin comparison were computed with a cross-validation approach, separately for each database.

In the case of the IDEM/SPREAD database, the uncalibrated scores were computed by doing all the possible comparisons between a session of a speaker and the other session of the same (for same-origin) or different (for different-origin) speaker. For each comparison, reference population model has been trained using all database content except the speaker (for same-origin) or the two speakers (for different-origin) used for the comparison: in this way, we avoided probable bias. The calibrated scores were computed by calibrating each single score, while the others have been used to train the LogReg procedure.

The above approach was not feasible for the IMPAVIDO database, due to its much larger number of speakers. In this case, a K-fold cross-validation has been used, in which the database content was split into K equi-sized groups of speakers. Each group was used to make all the possible comparisons among its members, while the remaining K-1 groups were used to train the model of the reference population for those comparisons. The calibrated scores were computed in a similar way.

All the results were calculated considering both the full dimensional problem, i.e., the four vowels together, and separately vowel-by-vowel, so that the behaviour of LogReg calibration was checked in both cases.

The fusion was performed for each couple of available emulators, for single-vowel and all-vowels conditions, fusing uncalibrated scores and fusing already calibrated scores.

5. *Results and discussion*

In the Table 1, the achieved Cllr values for uncalibrated scores, and for those calibrated with the different LogReg implementation, for each emulator, in all-vowel and single-vowel condition, for both the selected IDEM/SPREAD and IMPAVIDO database are showed.

In any case, it is confirmed that LogReg calibration always improves the Cllr. The best calibration variant, in the sense that leads to the lowest mean Cllr with the lowest standard deviation, is not very clear because it seems to be dependent on the emulator and the database. The best result has been achieved with the Bosaris toolkit with the eIDEM_LK emulator, using all the vowels, for the IMPAVIDO database.

Table 1 - *Results of calibration of single method (eIDEM.aeio means that all vowels have been used; eIDEM.a means that one vowel only has been used). Each cell contains the mean value of the log-likelihood-ratio cost (Cllr) and its standard deviation (in parentheses). "Raw" means "uncalibrated"*

	Selected IDEM/SPREAD database					*IMPAVIDO database*				
System id.	Raw	Focal	Morr.	Bosaris	Mnrfit	Raw	Focal	Morr.	Bosaris	Mnrfit
eIDEM .aeio	0,30 (0,1)	0,05 (0,02)	0,04 (0,01)	0,04 (0,02)	0,12 (0,03)	0,04 (0,02)	0,01 (0,01)	0,012 (0,001)	0,02 (0,04)	0,011 (0,009)
eIDEM .a	0,80 (0,1)	0,8 (0,1)	0,23 (0,04)	0,23 (0,04)	0,70 (0,1)	0,74 (0,04)	0,14 (0,02)	0,120 (0,009)	0,74 (0,04)	0,29 (0,02)
eIDEM _NOTH. aeio	0,98 (0,01)	0,04 (0,02)	0,035 (0,01)	0,04 (0,02)	0,12 (0,03)	1,000 (0,009)	0,01 (0,02)	0,006 (0,002)	0,03 (0,06)	0,04 (0,11)
eIDEM _NOTH.a	0,99 (0,01)	0,27 (0,03)	0,27 (0,03)	0,27 (0,03)	0,91 (0,06)	1,000 (0,009)	0,14 (0,01)	0,14 (0,01)	0,14 (0,01)	0,34 (0,02)
eIDEM _LK.aeio	15 (0,5)	0,11 (0,02)	0,11 (0,02)	0,11 (0,02)	0,52 (0,09)	4,6 (0,1)	0,1 (0,3)	0,020 (0,001)	0,004 (0,003)	0,009 (0,003)
eIDEM _LK.a	3,86 (0,09)	0,24 (0,04)	0,23 (0,04)	0,24 (0,04)	0,65 (0,07)	2,05 (0,05)	0,11 (0,01)	0,113 (0,008)	0,109 (0,009)	0,27 (0,02)
eSMART .aeio	3,6 (2,1)	0,29 (0,10)	0,3 (0,1)	0,29 (0,10)	0,80 (0,2)	3.1 (1.5)	0,15 (0,04)	0,146 (0,036)	0,15 (0,04)	0,33 (0,07)
eSMART .a	2,0 (1,2)	0,35 (0,15)	0,3 (0,1)	0,34 (0,14)	1,1 (1,0)	1.5 (0,5)	0,21 (0,03)	0,20 (0,03)	0,21 (0,03)	0,50 (0,06)

Another research question of this study was to verify whether it is more convenient, in terms of Cllr, to combine emulators, each working on a different vowel, than having a single emulator that works on all the vowels simultaneously. The answer, as shown in the Table 2, is that it seems to depend on the FSASR emulator and the database. Anyway, the results clearly show that for the eSMART emulator is always better to let it work on each vowel separately and then fusing the scores. This means that the current implementation of eSMART (inspired to Bove *et al.*, 2002), that uses all the 4 vowels together, can be improved.

The next question is to decide for the best emulator, in terms of Cllr. The answer is achieved by means of a paired samples t-test (1 tail), $\alpha = 5\,\%$ on the Cllr values, and it is depicted in the Table 3. Using the selected IDEM/SPREAD database the best Cllr is achieved with eIDEM_NOTH emulator (working on all the vowels) and Morrison's LogReg variant for calibration, while using the IMPAVIDO database

the best result is given by eIDEM_LK and Bosaris toolkit for calibration. Even in this experiment, the result achieved using IMPAVIDO database was far better.

Table 2 - *Experimental results about the fusion of the single-vowel systems. In each cell of the table, the best fusion strategy, and the achieved mean log-likelihood-ratio cost (Cllr) are indicated*

	Selected IDEM/SPREAD database	*IMPAVIDO database*
Focal		eSMART: fusing the RAW single-vowel systems (Cllr=0,6)
Morrison	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,04) eSMART: fusing the RAW single-vowel systems (Cllr=0,04)	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,017) eSMART: calibration of the single-vowel systems, then fusion (Cllr=0,06)
Bosaris	eIDEM_LK: fusing the RAW single-vowel systems (Cllr =0,07) eSMART: fusing the RAW single-vowel systems (Cllr =0,254	eSMART: fusing the RAW single-vowel systems (Cllr=0,05)
Mnrfit	eIDEM_LK: fusing the RAW single-vowel systems (Cllr=0,14) eSMART: fusing the RAW single-vowel systems (Cllr=0,30)	eSMART: calibration of the single-vowel systems, then fusion (Cllr=0,06)

Table 3 - *The performance of the single emulators. Log-likelihood-ratio cost (Cllr) vales are expressed as mean value and standard deviation (in parentheses)*

	Selected IDEM/SPREAD database	*IMPAVIDO database*
Focal	eIDEM Full dim.calibrated: Cllr= 0.05 (0.02) eIDEM_NOTH Full dim.calibrated: Cllr= 0.04 (0.02)	eIDEM Full dim.calibrated: Cllr=0,01 (0,01)
Morrison	**eIDEM_NOTH Full dim. calibrated: Cllr = 0.03 (0.01)**	eIDEM_NOTH Full dim. calibrated: Cllr=0.006 (0.002)
Bosaris	eIDEM Full dim.calibrated: Cllr= 0.04 (0.02) eIDEM_NOTH Full dim.calibrated: Cllr= 0.04 (0.02)	**eIDEMLK Full dim. calibrated: Cllr = 0.004 (0.003)**
Mnrfit	eIDEM Full dim.calibrated: Cllr= 0.12 (0.03) eIDEM_NOTH Full dim.calibrated Cllr= 0,12 (0.03)	eIDEM_LK Full dim. calibrated: Cllr=0,009 (0,003)

The last experiment concerned with the fusing process of two different emulators. This is necessary in order to try to get a better new system. The findings show that in a very few cases we found a fusion which was better than any other "single" emulator in terms of Cllr (paired samples t-test (1 tail), α =5 %). Conversely, the

improvement was always very small, as highlighted in the Table 4. Anyway, among the experimental results we obtained, also the best-fused system (i.e. the one highlighted in Table *4*) was not better than the best single system, i.e. the eIDEM_LK full dimensional with Bosaris calibration. (Cllr= 0.004 (0.003)).

Table 4 - *results about the fusion of emulators. Log-likelihood-ratio cost (Cllr) values are expressed as mean value and standard deviation (in parentheses)*

Database	*Calib. method*	*Best achieved fusion*	*...better than the best single system, which is*
selected IDEM	Focal	eIDEM Full dim. uncalib. + eIDEM_NOTH Full dim. uncalib.: Cllr= 0,03 (0,02)	eIDEM_NOTH Full dim calib.: Cllr= 0.04 (0.02)
IMPAVIDO	Morrison	eIDEM_NOTHFull dim. uncalib + eIDEM_LK with Calibrated fused vowels: Cllr= 0,005 (0,002)	eIDEM_NOTH Full dim. Calib.: Cllr=0.006 (0.002)

6. *Conclusions and further remarks*

In the present study, the performances of two FSASR approaches traditionally used in Italy, namely IDEM/SPREAD and SMART, were evaluated through emulations included in the IMPAVIDO software, using LogReg techniques applied to calibration and fusion.

For this purpose, tests were conducted using the default database of IDEM/SPREAD, which does not contain socio-linguistic information about speakers. Since the database of SMART is not accessible, similar tests were also conducted on the default database of IMPAVIDO, which, like that of IDEM/SPREAD, does not contain socio-linguistic information about speakers. Furthermore, the tests were conducted on the vowels /a, ɛ, i, ɔ/, not on the complete vowel system of the Italian language, to maintain the same limitation as the IDEM/SPREAD software.

Despite these limitations, the work presents significant strengths. As far as is known, fusion procedures have never been applied (nor published) to either IDEM/SPREAD or SMART, neither previously documented in literature, nor by the authors of the respective systems, nor by independent researchers. Furthermore, it does not seem that the two systems, with calibration and fusion applied, have been compared in this manner before. Additionally, there is also no previous indication of a software tool, similar to the one developed and employed by the authors, being available to conduct such a comparative study.

The first finding is that, according to the experimental hypothesis, the LogReg calibration has improved the Cllr in every circumstance, suggesting that also the accuracy of the tested FSASR approaches may be improved introducing calibration.

The second finding is that the tested FSASR approaches give more accurate results when the IMPAVIDO database is used. This may be due to the larger

number of speakers and the required number of vowel occurrences with respect to the selected IDEM/SPREAD database.

These preliminary results suggest that a more detailed study would be necessary in order to explore the behaviour of such approaches depending on other factors such as the numerosity of the feature sets to compare (i.e. the number of occurrences for each vowel) and possibly the noise conditions. Reducing the requirements on vowel quantity, at least for anonymous sample, can help to simulate those caseworks where very limited data are available: this is a quite simple operation. On the other hand, increasing that prerequisite, or even replicate the same experiment with introducing a sufficiently large number of noisy speech recordings, is a challenge task because **one** need to accurately label a very large number of vowels, which is very time-consuming.

Also, both the tested databases do not distinguish between different speaking sessions of a same speaker; neither have any other meta-data such as noise level. On the other hand, calibration can be influenced by the actual conditions in which the speech recordings have been produced. In the present study, we have implicitly assumed that the features values in both the tested databases where representative of typical casework conditions, where the signal-to-noise ratio (SNR) of the speech was above 10 dB, which is a threshold commonly accepted to extract spectro-acoustic measurements on vowels. Probably, a deeper study of the impact of calibration on the tested FSASR would add more details to the picture.

Another important conclusion may be drawn specifically for the eSMART emulator. As already mentioned, it is a component of the software IMPAVIDO which was inspired to the work of Bove *et al.* (2002), which in turn can be considered the basis for the modern SMART software. The achieved results show that the way in which it uses the vowel information is not effective: a better system can be achieved by using eSMART on each single vowel, and then use LogReg to fuse the four scores into a calibrated and fused LR. This result appears to be of considerable importance. In fact, it suggests that even the original SMART software might function more effectively than it presumably does today. In such a case, it could reasonably be suggested to conduct partial tests on each vowel separately from the others, and then obtain the final comparison result through the fusion of the individual partial results. A specific study on that software is therefore highly suggested.

The tested fused systems were sometimes able to outperform each single system that participated to fusion, but it was not able to perform the best single system we achieved. This may be explained by the fact that, despite the difference in the statistical approaches between eIDEM and eSMART, both systems work on the same features, which may be somehow redundant. From this point of view, a future direction of investigation might be to fuse a FSASR system based on a different kind of features, or, even better, to fuse a semi-automatic with fully automatic approach.

References

Aitken, C.G.G., Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, Wiley.

Bove, T., Giua, P.E., Forte, A. & Rossi, C. (2002). Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti. In *Statistica*, 62(3), 475-490.

Brümmer, N. & Du Preez, J. (2006). Application-independent evaluation of speaker detection. In *Computer Speech and Language*, 20(2-3), 230–275.

Brümmer, N., & Villiers, E.D. (2013). The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF. ArXiv, abs/1304.2865.

Brümmer, N., Swart, A. & van Leeuwen, D. (2014). A comparison of linear and nonlinear calibrations for speaker recognition. In *Odyssey 2014: The Speaker and language Recognition Workshop*, Tokyo, Japan, 1-5- November, 14–18.

Cumani S., Laface, P. (2019). Tied Normal Variance–Mean Mixtures for Linear Score Calibration. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 12-17 May, 6121-6125.

Falcone, M., Paoloni, A., De Sario, N., Saverione, B. (1992). IDEM: un sistema per l'analisi e la rappresentazione vocale. In *Associazione Italiana di Acustica, XX Congresso Nazionale*, Rome, 20-24 Aprile, 417-422.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M & Ortega-Garcia, J., (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. In *Computer Speech and Language*, 20(2-3 SPEC. ISS.), 331–355.

Jessen, M. (2008). Forensic phonetics. In *Language and Linguistics Compass*, 2, 671–711.

Jona Lasinio, G., C. Rossi & T. Bove (2004). The Speaker Recognition problem. In *Proceedings of the XLII Scientific Meeting of the Italian Statistical Society (SIS)*, Palermo, 20-22 June, 429–440.

Kersta, L.G. (1962). Voiceprint identification. In *Nature*, 196, 1253–1257.

Morrison, G.S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). In *Speech Communication*, 53(2), 242–256.

Morrison, G.S., (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. In *Australian Journal of Forensic Sciences*, 45(2), 173-197.

Morrison, G.S., Enzinger M., Phil, E., Zhang, C. (2017). Forensic Speech Science 2017. 183–197. In Freckelton I., Selby H. (a cura di), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters, 1-140.

Plchot O., P. Matějka, Silnova, A., Novotny, O., Diez, M., Rohdin, J., Glembek, O., *et al.* (2017). Analysis and description of ABC submission to NIST SRE 2016. In *Interspeech Proceedings*. Stockholm, Sweden. 20-24 August, 1348-1352.

Ramos D., Krish R.P., Fierrez J., Meuwly D. (2017). From Biometric Scores to Forensic Likelihood Ratios. In Tistarelli, M., Champod, C. (a cura di), *Handbook of Biometrics for Forensic Science. Advances in Computer Vision and Pattern Recognition*. Berlin: Springer, 305-327.

Ramos, D. & Gonzalez-Rodriguez, J. (2013). Reliable support: Measuring calibration of likelihood ratios. In *Forensic Science International*, 230(1–3), 156–169.

Romito, L., Bove, T., Delfino, S., Rossi, C., Jona-Lasinio, G.J. (2009). Specifiche linguistiche del database utilizzato per lo Speaker Recognition in S.M.A.R.T. In Romito, L., Lio, R., Galatà, V. (a cura di), *La Fonetica Sperimentale: metodo e applicazioni*, Torriana: EDK Editore, 632–640.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor and Francis.

Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. In *Computer Speech and Language*, 20(2-3 SPEC. ISS.), 159–191.

Sigona, F., Grimaldi M. (2015). Il riconoscimento del parlante in ambito forense: uno studio indipendente sul software IDEM/SPREAD in uso ai Carabinieri. In *Sicurezza e Giustizia*, IV, 16–21.

Sigona, F., Grimaldi M. (2017). Tools for Forensic Speaker Recognition. In Orletti, F., Mariottini, L. (a cura di), Forensic Communication: Theory and Practice, Cambridge: Cambridge Scholars Publishing, 127-150.

Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J., & Nash, E. (1972). Experiment on Voice Identification. In *Journal of the Acoustical Society of America*, 51(6B), 2030–2043.

van Leeuwen, D.A., Brümmer, N. (2007). An introduction to application independent evaluation of speaker recognition systems. In Müller, C. (a cura di), *Speaker Classification I: Fundamentals, Features, and Methods*., Heidelberg: Springer-Verlag, 330–353.

van Leeuwen, D.A., Brümmer, N. (2013). The distribution of calibrated likelihood-ratios in speaker recognition. In *Interspeech Proceedings* (ISCA), Lyon, France, 25-29 August, 1619–1623, 2013.

DOI: 10.17469/O2111AISV000026

FRANCESCO SIGONA, GIUSEPPE VITOLO, MIRKO GRIMALDI

Forensic Automatic Speaker Recognition based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network

In the field of automatic forensic voice comparison (FVC), the use of neural networks models in the processing chain is increasingly frequent. Recently, Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) demonstrated high discrimination and accuracy in the non-forensic speaker verification task. In this contribution, after illustrating the fundamental differences between the forensic and non-forensic tasks of speaker verification – from a linguistic, logical and methodological perspective – the performances of a software implementation of automatic FVC based on ECAPA-TDNN, at different simulated operating conditions (noise level, net speech duration), are verified. Preliminary results confirm excellent performance in non-critical operating conditions. The hypotheses on the performance trend are also preliminarily confirmed: as the duration of the speech samples increases, and the noise level decreases, the evaluation metrics improve at a rate that depends on the combination of these two factors.

Keywords: Forensic linguistics, voice comparison, speaker verification, speaker recognition.

1. *Introduction*

Automatic Speaker Recognition (ASR), i.e. the task to verify if the characteristics of the voice sample of a test speaker match the enrolled model of another speaker, finds numerous applications, mainly where some kind of user authentication is required. Therefore, ASR is employed to grant authorized access to info-telematic services, or personal devices (for example, smartphones), or physical environments with restricted access, and so on, in the most varied application sectors (such as banking-financial and healthcare: Sigona, 2018).

On the other hand, speaker recognition also plays an extremely important role in the forensic field: Forensic Speaker Recognition (FSR, or Forensic Voice Comparison, FVC) makes it possible to establish from a probabilistic, qualitative and quantitative point of view, to what extent a recorded voice sample of unknown identity (the questioned-speaker recording) can be attributed to a particular speaker of known identity (e.g. a suspect speaker), according to some methodological requirements.

It is not surprising that ASR technologies are often used to build FASR (Forensic Automatic Speaker Recognition) systems (Slavíček, Swart, Klčo & Brümmer *et al.*, 2021, 2021; van der Vloed, Kelly & Alexander, 2020; Jessen, Meir & Solewicz,

2019): however, including such technologies in the forensic application is not straightforward, due to the delicacy of the matter and the consequences that the outcome of a vocal comparison can have on the fate of a judicial proceeding.

In this work, the fundamental differences between the forensic and non-forensic tasks of speaker verification, from a linguistic, logical and methodological perspective, are highlighted (cf. § 2.)

This contribution also presents one of the first implementations, with relative experimentation, of a FASR system based on Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN). This is a very recent neural network model which, for non-forensic applications, has already shown better performance in terms of recognition error rates, compared to previous models (Desplanques, Thienpondt & Demuynck, 2020). In § 3 the FVC based on ECAPA-TDNN is described, from a typical FVC processing chain to the actual implementation. In § 4, the experimental design used to achieve the system performance in different case-work operating conditions is illustrated, and in § 5 the corresponding results are presented. Finally, in § 6 the results are discussed, while conclusions are drawn in § 7.

2. *Forensic and non-forensic speaker verification tasks.*

Multiple ASR techniques have been developed over the years: first they model appropriate features of the voice sample of the user to be authenticated, and then compare this model with one or more pre-enrolled speaker models, usually stored in the recognition system. These techniques include the Joint Factor Analysis (Kenny, 2005; Kenny, Boulianne, Ouellet & Dumouchel *et al.*, 2007; N. Dehak, Kenny, R. Dehak, Glembek, Dumouchel, Burget, Hubeika & Castaldof, 2009), the i-vector-based paradigms (N. Dehak, Kenny, R. Dehak, Dumouchel & Ouellet, 2011), up to the more recent introduction of Deep Neural Networks (DNN) in some place in the processing chain (Lei, Scheffer, Ferrer, & Mclaren, 2014, among the first attempts) or as an end-to-end approach (Snyder, Ghahremani, Povey, Garcia-Romero, Carmiel & Khudanpur, 2016). The result of an ASR is generally a numerical value (score) which in most cases expresses little more than a similarity between the input and the enrolled model(s). The outcome of the procedure is binary (authenticated/rejected) and is determined by comparing the score against a threshold value, suitably designed to meet certain requirements in terms of probability of decision error (false authentication of a foreign speech sample, false rejection of a legitimate speech sample). These requirements are in turn chosen based on the best acceptable compromise between the two error rates, according to the addressed application.

The voice comparison in the forensic application is a conceptually very different thing, in which the "general-purpose" ASR technologies can certainly be exploited, but this constitutes only one of the building blocks of the system. The FVC is a particularly dedicated field of application: in this case the need for the comparison methodology to be based on quantitative measurements lead to acceptable error

rates, and above all to be reproducible and validated in the scientific community has been repeatedly highlighted (Enzinger, Morrison & Ochoa, 2016; Morrison, 2014). The most established methodology in the literature is based on a Bayesian approach recently also recommended by the European Network of Forensic Science Institutes (Drygajlo, Jessen, Gfroerer, Wagner, Vermeulen & Niemi, 2016). In this setting, the FVC is not configured as a decision test and therefore the outcome is not binary (identified/rejected). Rather, it is expressed in the form of a Likelihood Ratio (LR; Morrison, 2009; Morrison, 2011) between the prosecution hypothesis (of identification, in which the degree of similarity between the two test samples prevails) and the defence hypothesis (non-identification, in which the degree of typicality of the items prevails, i.e. the similarity with a set of samples taken from a suitable "reference population").

There are many factors that can influence the performance of an ASR system. With the same choice of algorithms and related parameters, a decisive role is played by the operating conditions of the comparison, case by case. These conditions include:

(1) the effect of the transmission/recording channel of the vocal sample, and therefore of alteration of the frequency spectrum of the voice.
(2) the level and type of noise present in the recordings.
(3) the net duration of the speech inside the recorded samples.

In the case of FVC, an additional condition must be taken in account:

(4) socio-linguistic factors such as gender and the linguistic variety of the speakers, which also concern the logical-methodological correctness of the comparison (in light of which the results of the comparison and the system performance metrics are to be interpreted.).

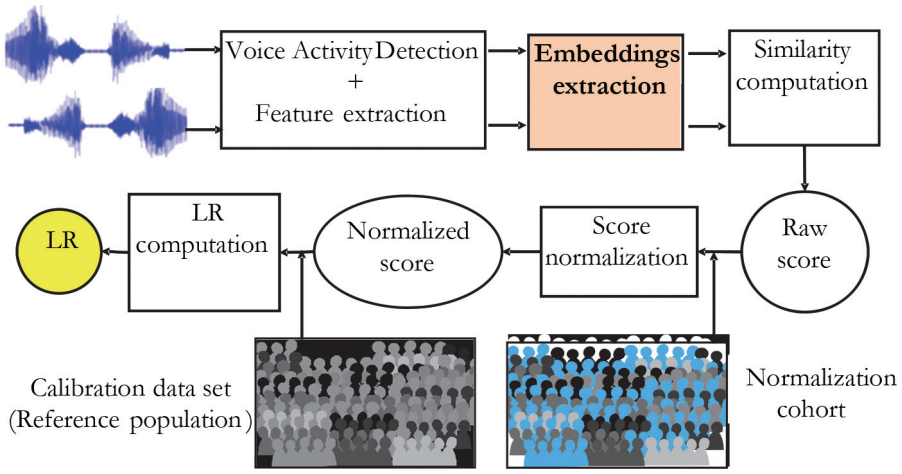
The linguistic variety of the spoken samples must be assessed to correctly formulate the defence hypothesis: accordingly, the questioned voice sample has been produced by a different speaker than the one investigated. It is immediate to understand that this "generic" different speaker cannot be one speaking a different language or variety but must be able to speak the same variety of the anonymous speaker as fluently as he does. Therefore, unlike the non-forensic ASR, in a FVC it is essential to have a so-called "reference population", which reflects not only the recording conditions of the case at hand, but also the linguistic variety. The selection of a suitable "reference population", in terms of recording conditions, linguistic variety, speaker gender and, where possible, also speaking style, is a key factor that can dramatically influence the outcome of FVCs.

3. *Typical processing chain for the FASR task.*

Figure 1 shows a typical sequence of steps for a FASR/FVC. The input of the system usually consists of vocal samples in which only the voice of the speaker of interest is present. Therefore, in the case of an interception of a conversation between two (or more) speakers, the forensic expert usually takes care of segmenting and

concatenating only the parts of the sample containing the requested voice, before introducing this data into the system. Each of the two resulting vocal samples to be compared are subsequently converted into a sequence of a variable number of fixed-length feature vectors, relating only to the spoken part (excluding any relevant pauses still present). These sequences constitute the input for the extraction of the vectors of embeddings, one vector for each of the two voice samples. The two vectors of embeddings are input into a similarity function to get a raw similarity score. A score normalization technique can also be used to mitigate the effect of mismatch between the recording conditions of the speech samples to be compared. This usually requires a normalization cohort, i.e. a dataset of speech recordings in various recording conditions, including those of the actual comparison. The last step is a score-to-forensic likelihood ratio conversion, also known as "calibration".

Figure 1 - *Schematic of a typical processing chain of a Forensic Voice Comparison system, from the input voice samples to the finale likelihood Ratio calculation*



3.1 Implementation of the processing chain

We have implemented the reference scheme illustrated in Fig. 1, as follows:

a. Input audio samples: each voice sample only contains voice segments, with pauses no longer than 100 ms.
b. Feature extraction and embeddings calculations: these steps are in charge to the pre-trained ECAPA-TDNN model, that will be presented in § 3.2.
c. Similarity computation: a raw comparison score is calculated by means of the well-known cosine similarity scoring function between the two input vectors of embeddings w_1 and w_2:

$$(1) \qquad score(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|}$$

d. Score normalization: Symmetric Normalization (S-norm) is used (Shum, N. Dehak, R. Dehak & Glass, 2010). The questioned-condition vector of embeddings

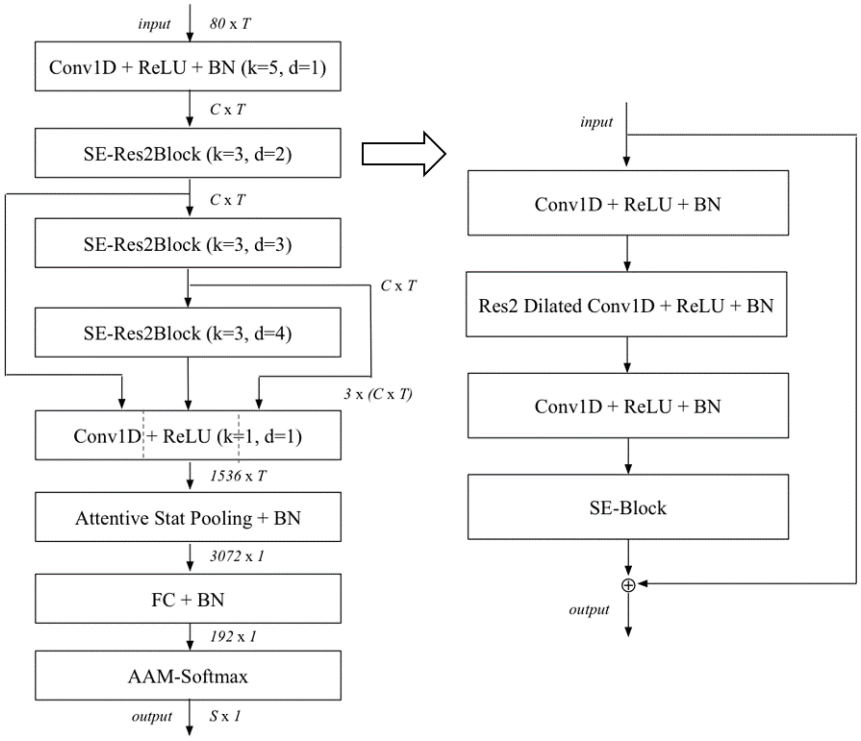
is compared with each vector of embeddings of the normalization cohort, producing a set of scores, from which the normalization statistics (mean and standard deviation) are calculated. The suspect-condition vector of embeddings is also compared with each vector embeddings of the cohort, producing a second set of scores, and a second pair of normalization statistics (a second mean and a second standard deviation). Finally, the final normalized score is obtained by normalizing the score with both the statistics sets (subtracting the mean and dividing by the standard deviation), and then taking the average of the two resulting values. In the current work, the normalization cohort is selected as a portion of whole dataset, while the remaining portion is used as reference population.

e. Score-to-forensic-likelihood-ratio conversion (calibration). For each speaker in the reference population, at least one speech sample in the questioned-speaker recording condition and one sample in the known-speaker recording condition must be available. By this way, it is possible to compute the set of the same-speaker scores (i.e. the scores between the two samples of each speaker), and the set of the different-speakers scores (i.e. the scores between the actual questioned-speaker sample and each sample of the reference population in the known-speaker recording conditions). The two set of scores are used to calculate respectively the same-speaker scores distribution and the different-speakers scores distribution, by means of a Kernel Density Estimate (Parzen, 1962). The same-speaker density evaluated at the actual score value has the meaning of a likelihood of the prosecution hypothesis, while the different-speaker density evaluated at the actual score value has the meaning of a likelihood estimate of the defense hypothesis. The LR is achieved as the ratio between these two likelihood values.

3.2 The ECAPA-TDNN model: a short overview.

According to Desplanques *et al.* (2020), ECAPA-TDNN includes multiple enhancements to the baseline TDNN-based x-vector architecture which applies statistics pooling to project variable-length utterances into fixed-length speaker characterizing embeddings. Such enhancements include restructuring the frame layers into 1-dimensional Res2Net modules with impactful skip connections, also introducing Squeeze-and-Excitation (SE) blocks in these modules to explicitly model channel interdependencies. Moreover, hierarchical features from different network layers are aggregated and propagated, and the statistics pooling module has been improved with channel-dependent frame attention, to make able the network to focus on different subsets of frames during each channel's statistics estimation.

Figure 2 - *Left diagram: network topology of the ECAPA-TDNN. k is the kernel size and d is the dilation spacing of the Conv1D layers or SE-Res2Blocks. C (=1024) and T are the channel and temporal dimension of the intermediate feature-maps respectively. S is the number of training speakers. BN stands for Batch Normalization (Ioffe and Szegedy, 2015). Right diagram: the SE-Res2Block of the ECAPA-TDNN architecture. The standard Conv1D layers have a kernel size of 1. The central Res2Net Conv1D with scale dimension s = 8 expands the temporal context through kernel size k and dilation spacing d. Adapted from Desplanques et al. (2020)*



In the present work, we used the ECAPA-TDNN model available on the Hugging Face platform[1], which was pre-trained using Voxceleb1 + Voxceleb2 data sets (Nagrani, Chung, Xie & Zisserman, 2020) and Additive Margin Softmax Loss. The network model has been implemented within the SpeechBrain toolkit (Ravanelli, Parcollet, Plantinga, Rouhe, Cornell, Lugosch, Subakan, Dawalatabad, Heba, Zhong, Chou, Yeh, Fu, Liao, Rastorgueva, Grondin, Aris, Na, Gao, De Mori, & Bengio, 2021), a well-known Open-Source Conversational AI Toolkit. The model is configured to work with 80 log Mel filter banks energies as frame level input features, 1024 channels in the convolutional layers, dimension of the bottleneck in the SE-Block and attention module set to 128, scale dimension *s* in the Res2Block

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb.

(Gao, Cheng, Zhao, Zhang, Yang & Torr, 2019) set to 8, and 192 nodes in the final fully-connected layer. The model has more than 20 million trainable parameters.

In the context of the present work, the model is essentially used to transform a vocal sample to its fixed-length vector of embeddings, for further processing.

3.3 Evaluation metrics

The performances of a FSR system can be represented by means of graphics (such as Tippet plots: Morrison, 2010), and numerical metrics, such as the Equal Error Rate (EER) and, even better, the Log-Likelihood Ratio Cost (Cllr).

The C_{llr} summarizes the overall system quality and have the following formulation (Brümmer and du Preez, 2006; van Leeuwen and Brümmer, 2007; Morrison, 2011):

$$(2) \qquad C_{llr} = \frac{1}{2}\left[\frac{1}{N_{ss}}\sum_{i=1}^{N_{ss}} \log_2\left(1+\frac{1}{LR_{ss_i}}\right) + \frac{1}{N_{ds}}\sum_{j=1}^{N_{ds}} \log_2\left(1+LR_{ds_j}\right)\right]$$

In the above formulation, LR_{ss} are the likelihood ratio values in the case of same-speaker comparison, while LR_{ds} are likelihood ratio values in the case of different-speakers comparisons. Since LR_{ss} values much greater than 1 better support the same-speaker hypothesis, and LR_{ds} values much less than 1 better support the different-speakers hypothesis, smaller values of C_{llr} indicate better performances. On the contrary, a system which provides no useful information and always responds with a likelihood ratio of 1 will result in a C_{llr} value of 1.

The EER is another largely used metric to evaluate the discriminating power of the system. Likelihood ratios test values can be combined with some prior odds to achieve posterior odds; in turn, posterior odd can be matched against a threshold to classify a test comparison as same-speaker (prosecutor hypothesis) or different-speakers (defense hypothesis). By this way, false "identification" and false "rejection" error rate can be computed as proportion of wrong classifications. Adjusting the priors and the threshold in such a way that the two error rates are equals, the resulting error rate is called the EER. This metric can be calculated using the Receiver Operator Characteristic Convex Hull method (Brümmer and de Villiers, 2013).

In this work, EER and C_{llr} are used to assess the performance of the tested FVC system.

4. *Experimental design*

In this pilot experiment, we wanted to test the pre-trained ECAPA-TDNN model as previously described, carrying out batteries of comparison tests between speech recordings in Italian. As far as we know, structured databases of Italian voice recordings with an adequate number of speakers, different sociolinguistic profiles and different speaking styles are not available for the validation of FVC systems. For this reason, a test dataset was built by the authors, collecting extracts of audiobook narratives. A dataset was created with 127 audiobook male narrators, collecting two recording excerpts for each speaker, lasting approximately three minutes each,

without any particular selection criteria, with the exception of the total absence of background melodies. Considering the goal of our work, this dataset can still be useful for the purposes of the study, and it can constitute the basis for a more structured and extensive activity of collecting vocal samples that can be used as a reference population.

Starting from this dataset, signal filtering, cutting and noise contamination has been used to simulate telephone-quality speech recordings, at different duration and different Signal-to-Noise Ratio (SNR) with Additive White Gaussian Noise (AWGN) and babble noise. Noise contamination at a given SNR value has been achieved by the FaNT tool (Filtering and Noise adding Tool)[2]. The whole information about the experimental conditions is summarized in Table 1. Different channels conditions and different types of noise will be experimented in future works.

Once the questioned- and known-speaker conditions have been chosen among those envisaged for the various tests, the questioned-speaker condition specifications are used to adapt all first recordings of the speakers in the dataset, and the known-speaker condition specifications are used to adapt the second recordings of the same speakers. This allows to build couples of recordings that can simulate actual caseworks in the selected operating conditions, in which same-speaker comparisons (when both the recordings belong to a same speaker) as well as differ-speakers comparisons (when the two recordings belong to different speakers) occur.

According to a one-two-speakers-out cross-validation paradigm, the pairs of recordings are extracted one at a time, and for each pair a new reference population is considered consisting of all the recordings that were not produced by the extracted speakers. By this way, for each couple, it is possible to compute an unbiased LR value, following the procedure illustrated in the **§** 3. The set of LR values is finally used to compute the evaluation metrics.

This process is reiterated for each operating conditions, to achieve evaluation metrics for each one.

Table 1 - *Summary of the parameters considered in the experimental design*

Spectral feature computation	80 mel filter bank, in charge to pre-trained ECAPA-TDNN models
Embeddings extraction	based on ECAPA-TDNN (192 embeddings)
Similarity	cosine distance
Score normalization	Adapted Symmetric Norm (AS-Norm)
Reference population	127 Italian male adults (audiobook narrators)
Channel emulation	filtering in the telephone bandwidth (300-3400 Hz)
Noise contamination	Additive White Gaussian Noise (AWGN); Babble noise; SNR levels: 6, 12, 18, 24 dB
Questioned- recording duration	5, 10, 20, 30, 40, 50, 60 s
Suspect-recording duration	10, 20, 30, 40, 50, 60 s
Metric for evaluation:	Cllr (Log-likelihood ratio cost), EER (Equal Error Rate)

[2] http://dnt.kr.hs-niederrhein.de/indexbd2f.html.

5. *Experimental results.*

Figure 3 shows the EER and C_{llr} trends, where the audio were filtered in the telephone band (300-3400 Hz), the known-speaker recordings have 24 dB SNR and different durations, while questioned-speaker recordings have various SNR values and various duration. All the audio samples we contaminated by AWGN. Figure 4 shows similar results achieved when all the audio samples we contaminated by a babble noise.

Figure 3 - *EER (upper plots) and C_{llr} (lower plots) as a function of the duration of the questioned-speaker recordings (horizontal axes), at different SNR values of the questioned-speaker recordings (colored lines) and different durations of the known-speaker recordings (from 10 s to 60 s, at the top of each plot). The known-speaker recordings have 24 dB SNR. All recordings have 300-3400 Hz frequency bands. Noise contamination with Additive White Gaussian Noise (AWGN)*

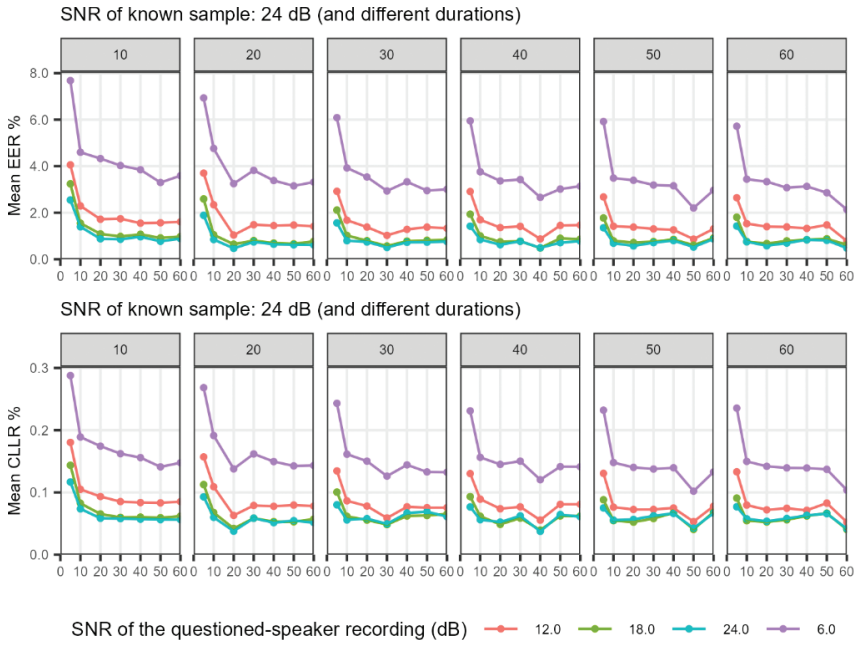
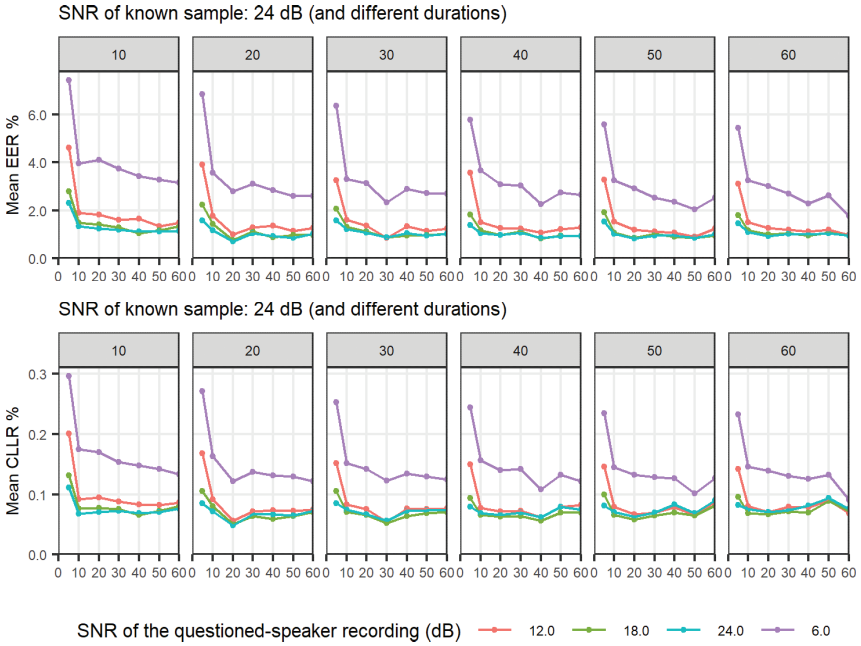


Figure 4 - *The same as Figure 3, but audio samples were contaminated by babble noise*



6. *Discussion*

The tests conducted and the results represented in Figures 3-4, although obtained with an unstructured dataset of speakers, not informed about the sociolinguistic conditions of the subjects, offer numerous insights for the forensic application of the tested system.

a) Duration of the suspect recording. Usually, suspect recording is carried out with a controlled procedure to obtain high quality audio, at most adapted, or adaptable in post-processing, to the recording channel of the questioned recording. In the tests carried out here, this situation can be traced back to the conditions of high SNR and long duration of the suspect recording.

b) Duration of questioned-speaker recording. There is a tendency to improve EER and C_{llr} as the duration of the questioned-speaker recording increases, going from 5 seconds to 60 seconds. This is an expected behaviour, because vocal samples that are too short may express only a small part of the characteristics of the speaker's voice, and therefore may lead to lower accuracy levels. As the net duration of the speech sample increases, so does the variety of these characteristics exhibited, and thus the system becomes able to build a more accurate model of the speaker's voice.

c) Effect of the loudness. The rate at which system accuracy improves as the voice sample length of the anonymous voice increases depends on the loudness of the signal (and, to some extent, on the duration of the known-speaker recording).

Indeed, when the anonymous sample has high SNRs, the accuracy improvement occurs suddenly, as early as 10 seconds, or at 20 seconds when the duration of the known sample is only 10 seconds. In general, increasing the duration of the known sample tends to improve performance when the conditions of the anonymous sample are not optimal (low duration, low SNR). As the SNR of the anonymous sample increases, it becomes less and less important to have high durations of the same sample. It should also be noted that the Cllr values remain below 0.1 already with values of 12 dB, for all the durations of the known sample. In particular, considering that the known sample is very often the result of a special recording, carried out in good recording conditions and without time limits, it can be stated that, except in the case of very noisy anonymous samples, the Cllr values settle comfortably below 0.05 in many operating conditions.

d) The system seems to perform quite similarly in both AWGN and babble noise conditions, at least in the scope of the tests conducted.

7. *Conclusions*

Our study provided preliminary results of the application of the ECAPA-TDNN model in the field of forensic voice comparison. Specifying "forensic" in the voice comparison task indicates that the speaker recognition procedure rises to a higher level of complexity. In addition to the similarity between the two voice recordings compared, their typicality, with respect to a reference population, must be taken into account, applying the Likelihodd Ratio Framework. The variety of conditions that can arise in a real case is such that it generally makes the task of testing an FVC system in all conditions very difficult. For this reason, FVC systems are most frequently tested under a few conditions, which are assumed to be more or less representative of one possible condition among many. One of the major difficulties in this sense is the lack, or in any case the difficulty, of building structured databases from a sociolinguistic point of view and of other factors susceptible to altering the speaking style. These difficulties also exist in particular for the Italian area. To overcome this problem, at least to a minimal extent, a dataset of speakers was specifically collected for the study presented here, without adopting particular selection criteria, and devoid of sociolinguistic information. Such a dataset, despite all the limitations already highlighted, still proved functional in highlighting the performance of the tested model, which is the main purpose of this work.

The experimentation carried out shows that a FVC based on the representation of voice utterances by means of embeddings calculated with ECAPA-TDNN can provide excellent levels of accuracy in the tested operating conditions. The results achieved encourage us to continue experimenting by introducing various other types of noise (e.g. babble, car, etc.) and with other reference populations, distinguished by gender of speakers and different linguistic varieties.

References

Brümmer, N. & Du Preez, J.A. (2006). Application-independent evaluation of speaker detection. In *Computer Speech & Language*, 20(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N.& de Villiers, E. (2013). The BOSARIS toolkit: theory, algorithms and code for surviving the new DCF. arXiv:1304.2865

Dehak, N., Kenny, P. Dehak, R., Glembek, O. Dumouchel, P., Burget, L., Hubeika, V., & Castaldo, F. (2009). Support vector machines and joint factor analysis for speaker verification. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (ICASSP'09), 4237–4240.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P. & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proc. Interspeech,* Brigthon, U.K.

Dehak, N., Kenny, P., Dehak, R., DumouchelP. & Ouellet, P. (2011). Front-end factor analysis for speaker verification. In *IEEE Trans. Audio Speech Lang. Process*. 19(2011) 788–798.

Desplanques, B., Thienpondt, J. & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, 3830-3834, doi: 10.21437/Interspeech.2020-2650.

Gao S., Cheng, M.M., Zhao K., Zhang, X., Yang, M.-H & Torr, P.H.S. (2019). Res2Net: A new multi-scale backbone architecture. In *IEEE TPAMI*, 2019.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2016). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises. *https://enfsi.eu/*

Enzinger, E., Morrison, G.S. & Ochoa F. (2106). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic voice-comparison case. In *Science & Justice*, 56, 1(2016): 42–57. DOI:10.1016/j.scijus.2015.06.005.

Jessen, M.E., Meir, G., & Solewicz, Y.A. (2019). Evaluation of Nuance Forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). In *Speech Communication*, 110, 101–107. https://doi.org/10.1016/j.specom.2019.04.006

Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. In*Tech. Rep. CRIM-06/08-13,* CRIM, Montreal, Quebec, Canada, 2005.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigen channels in speaker recognition. In *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 4, pp. 1435–1447, 2007.

Van Leeuwen, D.A., & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In *Lecture Notes in Computer Science*, 330–353, Springer Science+Business Media. https://doi.org/10.1007/978-3-540-74200-5_19

Lei, Y., Scheffer, N., Ferrer, L.& McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 1695–1699. doi: 10.1109/ICASSP.2014.6853887

Morrison, G.S. (2014). Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison. In*Science & Justice*, 54(2014): 245–256.

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. In *Science & Justice*, 51(3), 91–98. https://doi.org/10.1016/j.scijus.2011.03.002

Morrison, G.S. (2010). Forensic voice comparison. In Freckelton, I., Selby, H. (a cura di), *Expert Evidence*. Thomson Reuters, Sydney, Australia ch. 99. http://expert-evidence.forensic-voice-comparison.net

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. In *Science & Justice,* 49(2009): 298–308.

Nagrani, A., Chung, J.S., Xie, W. & Zisserman A. (2020). Voxceleb: Large-scale speaker verification in the wild. In *Comput. Speech Lang*., 60, 2020.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. In *Annals of Mathematical Statistics*, 33(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Ravanelli M., Parcollet T., Plantinga P., Rouhe A., Cornell S., Lugosch L., Subakan C., Dawalatabad N., Heba A, Zhong J., Chou JC., Yeh SL., Fu SW, Liao CF, Rastorgueva E., Grondin F., Aris W., Na H., Gao Y., De Mori R. & Bengio Y. (2021). In SpeechBrain: A general-purpose speech toolkit. *arXiv:2106.04624*.

Shum, S., Dehak, N., Dehak, R. & Glass, J. (2010). Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification. In *Odyssey*, 16. http://people.csail.mit.edu/sshum/papers/Odyssey10paper.pdf

Sigona, F. (2018). Voice Biometrics Technologies and Applications for Healthcare: an overview". In *Journal of interDisciplinaryREsearch Applied to Medicine (JDREAM)*, Vol.2, Issue 1, 2018, ESE - Salento University Publishing e-ISSN: 2532-7518, 5-15. http://siba-ese.unisalento.it/index.php/jdream; DOI: 10.1285/i25327518v2i1p5

Slavíček, J., Swart, A., Klčo, M., & Brümmer, N. (2021). The PhonexiaVoxCeleb Speaker Recognition Challenge 2021 SystemDescription. https://arxiv.org/abs/2109.02052

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y. & Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, 165–170. doi: 10.1109/SLT.2016.7846260

van der Vloed, D., Kelly, F. & Alexander, A. (2020). Exploring the Effects of Device Variability on Forensic Speaker Comparison Using VOCALISE and NFI-FRIDA, A Forensically Realistic Database. In P*roc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 402-407, doi: 10.21437/Odyssey.2020-57.

Autori

Anna Anastaseni – LFSAG, Università di Torino; GIPSA-lab, CNRS/ Université Grenoble-Alpes
anna.anastaseni@unito.it; https://orcid.org/0000-0003-4457-7154

Cecilia Andorno – Dipartimento di Studi Umanistici, Università di Torino, Italia
ceciliamaria.andorno@unito.it; https://orcid.org/0000-0003-4093-4425

Giuditta Avano – Università di Pisa, Italia
giuditta.avano@phd.unipi.it; https://orcid.org/0009-0009-0739-9804

Cinzia Avesani – [2]ISTC-CNR Padova, Italia
cinzia.avesani@cnr.it; https://orcid.org/0000-0001-9911-1189

Simona Bernardini – ABC BALBUZIE Ò, Italia
simonabernardini.psic@gmail.com

Andrea Bolioli – Independent researcher, Turin, Italy
andrea.bolioli@gmail.com

Sabrina Bonichini – Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, Italia
s.bonichini@unipd.it; https://orcid.org/0000-0003-0143-572X

Mirko Bonotto – Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, Italia
mirko.bonotto@studenti.unipd.it

Alessio Brutti – Fondazione Bruno Kessler, Italy
brutti@fbk.eu; https://orcid.org/0000-0003-4146-3071

Franca Cameriero – Universo Salute – Opera Don Uva (PZ), Italia
francacameriero@gmail.com

Raffaello Caserta – Department of Computer Engineering, University of Enna Kore, Italy
raffaello.caserta@unikore.it

Federica Cavicchio – CRIL, Università del Salento, Italia
federica.cavicchio@unisalento.it; https://orcid.org/0000-0002-2795-8396

Federica Chiari – Dipartimento di Neuroscienze, Università di Padova, Italia
federica.chiari@unipd.it

Virginia Chiti – Università degli Studi di Firenze, Italia
virginia.chiti16@gmail.com

Valentina Colonna – Departamento de Lingüística General y Teoría de la Literatura, Universidad de Granada, Spagna
valentinacolonna@ugr.es; https://orcid.org/0000-0002-5411-1596

Claudia Crocco – Department of Linguistics, Ghent University, Belgium
claudia.crocco@ugent.be; https://orcid.org/0000-0003-1099-956X

Sonia d'Apolito – Department of Humanities, University of Salento, Italy
sonia.dapolito@unisalento.it https://orcid.org/0009-0002-5593-5076

Rosa De Micco – Dept. of Advanced Medical and Surgical Sciences, University of Campania Luigi Vanvitelli
rosita.demicco@gmail.com; https://orcid.org/0000-0002-6397-8888

Bianca Maria De Paolis – Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Università di Torino, Italia
biancamaria.depaolis@unito.it; https://orcid.org/0000-0001-7725-9617

Gaetano Di Bello – Universo Salute – Opera Don Uva (PZ), Italia
gaetanodibello@yahoo.it

Simone Dipace – University of Bari "Aldo Moro", Bari, Italy
simodip9@gmail.com; https://orcid.org/0009-0002-8434-5251

Daniele Falavigna – Fondazione Bruno Kessler, Italy
falavi@fbk.eu; https://orcid.org/0000-0002-4844-5071

Claudio Ferrante – Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Italy
claudio.ferrante@mail.polimi.it

Alessandra Ferrera – FICLIT, University of Bologna, Italy
alessandra.ferrera3@studio.unibo.it; https://orcid.org/0009-0006-4540-0384

Maria Luisa Fiorella – Department of translational biomedicine and neuroscience, University of Bari "Aldo Moro", Bari, Italy
marialuisa.fiorella@uniba.it; https://orcid.org/0000-0001-9199-4132

Gloria Gagliardi – Alma Mater Studiorum – FICLIT, University of Bologna, Italy
gloria.gagliardi@unibo.it; https://orcid.org/0000-0001-5257-1540

Erika Ganassin – Dipartimento di Neuroscienze, Università di Padova, Italia
erika.ganassin@studenti.unipd.it

Vita Garrammone – Universo Salute – Opera Don Uva (PZ), Italia
vgarrammone@virgilio.it

Sylvain Gerber – GIPSA-lab, CNRS/Université Grenoble-Alpes
silvain.gerber@grenoble-inp.fr

Barbara Gili Fivela – Laboratory CRIL & DReaM, Department of Humanities, University of Salento, Lecce, Italy
barbara.gilifivela@unisalento.it; https://orcid.org/0000-0002-4694-8652

Giovanna Gison – Centro Medico Riabilitativo Pompei (NA), Italia
giovannagison@gmail.com

Mirko Grimaldi – Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL) – Dipartimento di Studi Umanistici, Università del Salento, Italia
mirko.grimaldi@unisalento.it; https://orcid.org/0000-0002-0940-3645

Kim A. Groothuis – Department of Linguistics, Ghent University, Belgium
kim.groothuis@ugent.be; https://orcid.org/0000-0002-9887-0690

Teresanna Illuzzi – University of Bari "Aldo Moro", Bari, Italy
teresannailluzzi@gmail.com; https://orcid.org/0009-0006-8105-4175

Syria Cira Imparato – Alma Mater Studiorum – Università di Bologna, Italia
syriacira.imparato@studio.unibo.it

Maria Izzo – Centro Medico Riabilitativo Pompei (NA), Italia
maria73izzo@gmail.com

Valentina Lavermicocca – Department of translational biomedicine and neuroscience, University of Bari "Aldo Moro", Bari, Italy
valentina.lavermicocca@uniba.it; https://orcid.org/0009-0002-7426-9670

Giovanni Leo – Department of Linguistics, Ghent University, Belgium
giovanni.leo@ugent.be; https://orcid.org/0000-0002-9033-1131

Olga Liguori – Centro Medico Riabilitativo Pompei (NA), Italia
olgaliguori68@tiscali.it

Federico Lo Iacono – Università di Torino e Università di Genova, Laboratorio di Fonetica Sperimentale "A. Genre" (LFSAG), Università di Torino, Italia
federico.loiacono@unito.it; https://orcid.org/0009-0007-1685-0623

Roberta Bianca Luzietti – Università di Pisa, Italy / Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa, Italy
roberta.luzietti@phd.unipi.it; robertabianca.luzietti@ilc.cnr.it; https://orcid.org/0000-0002-1825-0097

Marta Maffia – Dept. of Literary, Linguistics and Comparative Studies, University of Naples L'Orientale
mmaffia@unior.it; https://orcid.org/0000-0002-4913-374X

Giuseppe Magistro – Department of Linguistics, Ghent University, Belgium
giuseppe.magistro@ugent.be; https://orcid.org/0000-0002-0272-741X

Elena Martinelli – Università di Bologna, Italia
elena.martinelli12@unibo.it; https://orcid.org/0009-0007-4399-6951

Matilde Martino – Universo Salute – Opera Don Uva (PZ), Italia
matmart1@virgilio.com

Francesca Mori – Universo Salute – Opera Don Uva (PZ), Italia
francesca.mori@hotmail.it

Nicholas Nese – Dipartimento di Studi Umanistici, Università di Pavia, Italia
nicholas.nese@unipv.it; https://orcid.org/0000-0003-3234-2577

Incoronata Nolè – Universo Salute – Opera Don Uva (PZ), Italia
incoronatanole1@gmail.com

Ernesto Orsino – Centro Medico Riabilitativo Pompei (NA), Italia
e.orsino@inwind.it

Alessandro Panunzi – Università degli Studi di Firenze, Italia
alessandro.panunzi@unifi.it; https://orcid.org/0000-0001-5127-1036

Anna Chiara Pagliaro – Department of Humanities, University of Salento, Italy
annachiara.pagliaro@unisalento.it https://orcid.org/0009-0001-7332-7587

Giulia Perfetto – Centro Medico Riabilitativo Pompei (NA), Italia
giugiu25@gmail.com

Duccio Piccardi – Dipartimento di Studi Umanistici, Università degli Studi di Urbino Carlo Bo, Italy
duccio.piccardi@uniurb.it; https://orcid.org/0000-0002-4985-4360

Serge Pinto – CNRS, Aix-Marseille University, LPL, Aix-en-Provence, France
serge.pinto@univ-amu.fr; https://orcid.org/0000-0002-3375-4787

Caterina Pisciotta – Independent researcher, Padova, Italia
caterina.pisciotta@gmail.com

Eleonora Rodegher – Libera Università di Bolzano, Italy
eleonorarodegher97@gmail.com

Antonio Romano – Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Laboratorio di Fonetica Sperimentale "A. Genre" (LFSAG), Università di Torino, Italia
antonio.romano@unito.it; https://orcid.org/0000-0002-2987-3062; https://orcid.org/0000-0002-5029-4372

Valentina Saccone – Università degli Studi di Firenze, Italia
valentina.saccone@unifi.it; https://orcid.org/0000-0002-6222-6326

Vincenzo Sallustio – Phoniatrics and Communicative Disorders Rehabilitation Center, Dept. Rehabilitation, ASL Lecce, Italy
v.sallustio@alice.it

Tina Santarpia – Centro Medico Riabilitativo Pompei (NA), Italia
tina.83@hotmail.it

Loredana Schettino – Faculty of Education, Free University of Bozen-Bolzano. Interdepartmental Research Center Urban/Eco, University of Naples Federico II
lschettino@unibz.it; https://orcid.org/0000-0002-3788-3754

Diletta Sciancalepore – Department of translational biomedicine and neuroscience, University of Bari "Aldo Moro", Bari, Italy
log.sciacalepore@gmail.com; https://orcid.org/0009-0004-9948-2923

Vincenzo Scotti – Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Italy
vincenzo.scotti@polimi.it; https://orcid.org/0000-0002-1825-0097; https://orcid.org/0000-0002-8765-604X

Donata Sensale – Centro Medico Riabilitativo Pompei (NA), Italia
donata.sensale@gmail.com

Francesco Sigona – Laboratory CRIL (Centro di Ricerca Interdisciplinare sul Linguaggio) & DReaM, Department of Humanities, University of Salento, Lecce, Italy
francesco.sigona@unisalento.it; https://orcid.org/0000-0003-2939-0009

Isabella Laura Simone – School of Medicine, University "Aldo Moro" of Bari, Italy
isabellalaura.simone@uniba.it; isimone51@gmail.com; https://orcid.org/0000-0002-7429-3091

Marco Siniscalchi – Department of Computer Engineering, University of Enna Kore
marco.siniscalchi@unikore.it; https://orcid.org/0000-0002-0770-0507

Roberto Tedesco – Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Italy
roberto.tedesco@polimi.it3; https://orcid.org/0000-0002-2830-4247

Alessandro Tessitore – Dept. of Advanced Medical and Surgical Sciences, University of Campania Luigi Vanvitelli
alessandro.tessitore@unicampania.it; https://orcid.org/0000-0003-1740-8566

Mario Vayra – Università di Bologna/ ISTC-CNR Padova, Italia
mario.vayra@unibo.it; https://orcid.org/0000-0002-6198-2437

Pietro Vigorelli – Gruppo Anchise, Milan, Italy
pietro.vigorelli@gruppoanchise.it

Giuseppe Vitolo – Dipartimento di ingegneria dell 'innovazione, Università del Salento –Italy
giuseppevit@gmail.com

Fabiola Vullo – Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, Italia
fabiola.vullo@studenti.unipd.it

Elisa Zacheo – Independent researcher, Padova, Italia
logo.elisazacheo@gmail.com

Claudio Zmarich – CNR-ISTC & Università di Padova, Italia
claudio.zmarich@cnr.it; https://orcid.org/0000-0003-0384-4041

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazione fonologiche alle applicazioni tecnologiche. I testi sono selezionati attraverso un processo di revisione anonima fra pari e vengono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

Barbara Gili Fivela, professore associato presso l'Università del Salento, si occupa di fonetica e fonologia, con attenzione ai temi di variazione linguistica, ai contesti di apprendimento delle seconde lingue e delle patologie del parlato. Presidente dell'Associazione Italiana di Scienze della Voce a partire dal 2019 (il secondo mandato termina nel 2024), è autrice di oltre cento contributi su argomenti di fonetica e fonologia, pubblicati in volumi e riviste specialistiche nazionali ed internazionali.

Sonia d'Apolito è attualmente assegnista di ricerca presso l'Università del Salento. Le aree di interesse riguardano principalmente fonetica e fonologia, apprendimento di una seconda lingua e linguaggio patologico. Ha collaborato in due progetti PRIN occupandosi dell'analisi acustica e articolatoria del parlato di soggetti affetti da Parkinson. Autrice e Co-autrice di contributi di interesse nazionale e internazionale.

Anna Chiara Pagliaro, PhD, ha svolto attività di ricerca post-dottorato presso l'Università degli Studi di Napoli Federico II e l'Università del Salento. I suoi interessi scientifici riguardano la comunicazione parlata normofasica e patologica, la pragmatica, la cortesia linguistica e il linguaggio formulare. Tra le sue pubblicazioni più recenti si segnalano: Postponing construction: recurrence, specification and productivity, in J.-P. Colson (a cura di), Phraseology, Constructions and Translation: Corpus-based, Computational and Cultural Aspects, Louvain-la-Neuve, 2023; Phonological and Sociophonetic Information in Dysarthric Speech: A First Investigation on Italian, in R. Skarnitzl & J. Volín (a cura di), ICPhS 2023 (con B. Gili Fivela, S.I. d'Apolito).

Vincenzo Sallustio, Audiologo e Foniatra, è Dottore di ricerca in Fisiopatologia delle Comunicazioni Audioverbali. attualmente responsabile dell'U.O. di Foniatria e Riabilitazione dei Disturbi della Comunicazione della ASL Lecce. È coautore di numerose pubblicazioni di interesse internazionale. Ha collaborato in progetti di ricerca con la State University of New York at Buffalo, con Università del Salento e con l'Istituto Italiano di Tecnologia.

Maria Luisa Fiorella è Professore Ordinario in Audiologia e Foniatria, presso la Scuola di Medicina e Chirurgia dell'Università degli Studi di Bari. Responsabile Didattico del Corso di Laurea in Logopedia e Docente in diversi Corsi di Laurea della Scuola di Medicina. Responsabile del Programma Interdipartimentale sulla Riabilitazione delle patologie audiologiche, foniatriche e dei disturbi della deglutizione. Autrice di numerose pubblicazioni scientifiche.

Sabato Marco Siniscalchi è Professore Ordinario presso l'Università di Palermo. È membro eletto del consiglio direttivo dell'ISCA e del comitato tecnico IEEE per il linguaggio e la parola (2024–2026). Ricopre il ruolo di Senior Associate Editor della rivista IEEE Signal Processing Letters ed è stato Associate Editor di IEEE Transactions on Audio, Speech, and Language Processing. La sua ricerca si concentra sul deep learning.

AISV - Associazione Italiana Scienze della Voce
sito: www.aisv.it
email: aisv@aisv.it| redazione@aisv.it
ISBN: 978-88-97657-70-5

Edizione realizzata da
Officinaventuno
info@officinaventuno.com | sito: www.officinaventuno.com
via F.lli Bazzaro, 18 - 20128 Milano - Italy