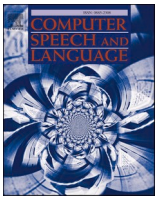
Computer Speech and Language 89 (2025) 101691

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



A computational analysis of transcribed speech of people living with dementia: The Anchise 2022 Corpus

Francesco Sigona [a,*], Daniele P. Radicioni [b], Barbara Gili Fivela [a], Davide Colla [c], Matteo Delsanto [b], Enrico Mensa [b], Andrea Bolioli [d], Pietro Vigorelli [e]

[a] Laboratory CRIL & DReaM, Department of Humanities, University of Salento, Lecce, Italy
[b] Department of Computer Science, University of Turin, Turin, Italy
[c] Department of Historical Studies, University of Turin, Turin, Italy
[d] Independent Researcher, Turin, Italy
[e] Gruppo Anchise, Milan, Italy

ARTICLE INFO

Keywords:
MMSE
Automatic speech and language analysis
NLP
Digital linguistic biomarkers
Emotion analysis
Perplexity
Naturalistic conversations
Enabling approach

ABSTRACT

Introduction: Automatic linguistic analysis can provide cost-effective, valuable clues to the diagnosis of cognitive difficulties and to therapeutic practice, and hence impact positively on well-being. In this work, we analyzed transcribed conversations between elderly individuals living with dementia and healthcare professionals. The material came from the Anchise 2022 Corpus, a large collection of transcripts of conversations in Italian recorded in naturalistic conditions. The aim of the work was to test the effectiveness of a number of automatic analyzes in finding correlations with the progression of dementia in individuals with cognitive decline as measured by the Mini-Mental State Examination (MMSE) score, which is the only psychometric-clinical information available on the participants in the conversations. Healthy controls (HC) were not considered in this study, nor does the corpus itself include HCs. The main innovation and strength of the work consists in the high ecological validity of the language analyzed (most of the literature to date concerns controlled language experiments); in the use of Italian (there is little corpora for Italian); in the size of the analyzed data (more than 200 conversations were considered); in the adoption of a wide range of NLP methods, that span from traditional morphosyntactic investigation to deep linguistic models for conducting analyzes such as through perplexity, sentiment (polarity) and emotions.

Methods: Analyzing real-world interactions not designed with computational analysis in mind, such as is the case of the Anchise Corpus, is particularly challenging. To achieve the research goals, a wide variety of tools were employed. These included traditional morphosyntactic analysis based on digital linguistic biomarkers (DLBs), transformer-based language models, sentiment and emotion analysis, and perplexity metrics. Analyzes were conducted both on the continuous range of MMSE values and on the severe/moderate/mild categorization suggested by AIFA (Italian Medicines Agency) guidelines, based on MMSE threshold values.

Results and discussion: Correlations between MMSE and individual DLBs were weak, up to 0.19 for positive, and -0.21 for negative correlation values. Nevertheless, some correlations were statistically significant and consistent with the literature, suggesting that people with a greater degree of impairment tend to show a reduced vocabulary, to have anomia, to adopt a more informal linguist register, and to display a simplified use of verbs, with a decrease in the use of participles,

* Corresponding author.
E-mail address: francesco.sigona@unisalento.it (F. Sigona).

https://doi.org/10.1016/j.csl.2024.101691
Received 30 November 2023; Received in revised form 27 June 2024; Accepted 20 July 2024
Available online 22 July 2024

gerunds, subjunctive moods, modal verbs, as well as a flattening in the use of the tenses towards the present to the detriment of the past. The -0.26 inverse correlation between perplexity and MMSE suggests that perplexity captures slightly more specific linguistic information, which can complement the MMSE scores. In the categorization tasks, the classifier based on DLBs achieved an F1 score of 0.79 for binary classification between SEVERE and MILD, and 0.61 for multi-label categorization. Sentiment and emotion analyzes showed inverse trends for joy while MMSE scores suggested that less impaired individuals were less joyful, or more "negative", than others. Considering the real-world context, this is consistent with the hypothesis of a gradual reduction in awareness in individuals affected by dementia. Finally, integrating various profiles of analysis has been proved to be effective in offering a wider picture of linguistic and communication deficits, as well as more precise data regarding the progression of dementia.

1. Introduction

Detecting language impairment has become increasingly important in identifying and diagnosing neurodegenerative disease. Language deficit can be observed in several neurodegenerative conditions, either as a preeminent symptom in early stages, as in Primary Progressive Aphasia, or in conjunction with other cognitive disorders, such as in Alzheimer's Disease (AD) (Boschi et al., 2017), as well as in individuals with Mild Cognitive Impairment (MCI) who later receive a diagnosis of Dementia (Mura et al., 2014; Karr et al., 2018). The analysis of connected speech produced by individuals with MCI related to AD and their cognitively healthy counterparts has revealed significant differences that primarily concern semantic impoverishment and reduced fluency (Filiou et al., 2020; Mueller et al., 2018). Consistently, in studies comparing individuals with AD who progress with those who do not, significant differences have been found in naming ability and semantic fluency (Kim et al., 2019; Vaughan et al., 2018).

Language difficulties pose a significant challenge for most dementia patients, particularly as the disease advances. The initial indicators of linguistic and communication impairments (LCIs) manifest themselves as difficulty in finding words, especially when it comes to identifying familiar individuals or objects. Incorrect and meaningless words replace the intended ones, and speech pauses become more frequent (Banovic et al., 2018). In the early stages of AD, for instance, language impairment primarily involves difficulties in word retrieval, diminished verbal fluency, and a breakdown at higher levels of written and spoken language. As AD progresses to moderate and severe phases, verbal fluency is greatly compromised, comprehension declines, and various types of paraphrases become prominent. In the most advanced stages of AD, speech often becomes limited to repetitive echoing (echolalia) and verbal stereotypes (see Ferris and Farlow, 2013; Soria Lopez et al., 2019).

Based on such evidence and given that the production of speech represents a task that closely reflects real-life situations and is crucial to daily functioning, the analysis of speech promises to be effective both in assessing cognitive abilities per se and as a barometer of disease progression. However, evaluating linguistic abilities using standard psychometric instruments is a time-consuming process and is susceptible to human bias. Recent studies evaluating LCI tests highlighted lack of standardization, normative data and criterion validity, as well as scant evidence attesting the reliability of those tools that had originally been developed for non-neurodegenerative LCIs (Krein et al., 2019). Further, it was observed that no tool considered the perspective of individuals living with dementia (ILWD) regarding the impact of LCIs on their daily life (Krein et al., 2019). In this work, linguistic abilities are evaluated by considering the speech produced by ILWD in a genuinely authentic relational communicative context.

The significance of automatic analysis of speech and language using Natural Language Processing (NLP) techniques for language and communication assessment has steadily increased, providing valuable insights into language proficiency and cognitive abilities, besides offering information on motor skills where the audio is available. There are many studies in the literature on NLP methods that set out to identify characteristics provided by text-based automatic analysis and correlate them with psychometric scores or use them to discriminate between healthy and unhealthy subjects (Vigo et al., 2022). However, at present, the quality and quantity of information is inadequate to provide recommendations regarding the choice of tasks, features, or algorithms that a clinician should organize or carry out (Gagliardi, 2023).

Also, most of these studies concern very controlled experiments, in which subjects carry out pre-established non-dialogical tasks which would introduce a certain level of structure to the speech output, producing semi-spontaneous language, such as picture description, reading or narrative tasks) (Prins and Bastiaanse, 2004).

Additionally, to date, the study of spontaneous oral productions of ILWD, in the context of interactive communication during everyday life, and therefore in conditions of high ecological validity, seems under researched. In a naturalistic setting, speakers do not perform an experimental task, but rather they are completely free of any research requirements, as they are free to express themselves. This is crucially different from what happens in controlled experiments. In this respect, conversations in naturalistic settings become a precious source of information, integrating the knowledge acquired through controlled experiments, and offering a more complete overall picture of LCIs.

An important source of data for this type of investigation is represented by the Anchise Corpus, a collection of manual transcripts of face-to-face conversations between ILWD, guests of nursing homes in various Italian locations, and operators who adopted the *ApproccioCapacitante* ® (Enabling Approach). The collected dialogues can be considered as characterized by high ecological validity if compared to productions elicited by means of reading tasks, picture description tasks or purely narrative tasks (Section 4.1). Despite some intrinsic limitations, including the unavailability of audio recordings of the conversations, the Anchise Corpus represents a formidable resource, not only because of the naturalness of the language, but also for the number of samples. Over 200 conversations,

involving as many ILWD subjects, distributed almost over the entire MMSE scale, represent a remarkable sample size given that in other works, which refer to more controlled trials, the sample size is considerably smaller (for the Italian, see De Stefano et al., 2021, which studied 47 unhealthy individuals; L. Calzà et al., 2021, with 48 unhealthy individuals; Dovetto et al., 2022, with about 20 unhealthy individuals. See also the very recent review by Yang et al., 2022).

The main goal of the investigations described in this paper is to test the effectiveness of a number of automatic analyzes of transcribed dialogues occurring in highly naturalistic contexts of Italian speakers affected by dementia at various stages (based on MMSE score), using the Anchise Corpus. As for the NLP analyzes, we discuss both the use of statistics regarding linguistic features (i.e. digital linguistic biomarkers, DLB) and the use of language models (LMs). Further, we focus our attention on lexical and morpho-syntactic features and variations in attitude, including sentiment and emotions. Based on this information we investigate if and to what extent we are able to find correlations between the speech transcripts and patients' MMSE scores.

This work expands on preliminary studies (Bolioli et al. 2020; Benvenuti et al. 2021; Sigona et al., 2023), in which the Anchise corpus was subjected to a more limited analysis, carried out with only a slightly smaller number of DLBs but with no deep language models.

The article is structured as follows: we first survey related work (Section 2) and analytically state the goals of the analyzes described in this paper (Section 3); we then illustrate materials and methods (Section 4): in particular, a statistical description of the main features of the analyzed corpus (Section 4.1) and the experiments carried out on the corpus along with the employed methods, based both on traditional NLP features and on the adoption of LMs (Section 4.2). We then report and discuss the results obtained through the array of experiments (Section 5), while the final Section provides a summary of the main contributions of the work and an outline of future work.

2. Related work

2.1. Tasks to elicit connected speech and speech resources

The availability of usable corpora of speech of people with cognitive impairments depends on the language. For example, the recent systematic review by Vigo et al. (2022), reports (see Table 2 of that article): 4 databases in English (including Talkbank-DementiaBank, by Becker et al., 1994, with 169 CE subjects), 4 in French, 4 in Spanish, 3 in Swedish and 3 in Turkish. However, the same review reports only 2 databases for Mandarin, 1 for Greek, 1 for Hungarian, and 1 for Italian (from the OPLON project: Beltrami et al., 2016, Beltrami et al., 2018; L. Calzà et al., 2021). The publication of two new corpora for Italian, such as the CIPP-ma and CIPP-mci (Dovetto et al., 2022), is only a partial contribution as it only includes a maximum of 40 patients and 40 health controls.

It should also be noted that even if languages refer to the same abstract structure and units, the way they select and realize them differs. For instance, all languages distinguish parts-of-speech (POS), but they differ regarding both the number and the kind of distinctions they make (Schachter and Shopen, 2007). English, for example, has articles while many Asian languages do not. Italian, on the other hand, like English has articles. However, it is a strongly inflected language, while English is not. These cross-language linguistic differences may regard all levels of the grammar (see, for example, Shopen, 2007) and have to be accounted for, as the analysis of datasets in different languages could offer different insights on the linguistic features of speech by ILWDs, e.g., as for some parts-of-speech. Also, though current technology is able to account for cross-language differences through transfer learning, the existence of language-specific corpora is undoubtedly a precious source of specific information that should be exploited whenever possible.

Thus, resources for Italian are significantly limited, very often they involve a small number of participants and have not been published. Accessibility is also an issue, since voice recordings of individuals are considered personal data (and hence subject to the principles of data protection), and European legislation does not allow voice recordings to be shared. The main accessible contributions of semi-spontaneous language regarding varieties of Italian are the corpora mentioned above, which have been collected by adopting quite standardized tasks.

In general, one commonly used method for obtaining speech samples requires patients to describe a scene, such as in the "Cookie Theft" picture (Goodglass et al., 1983). Picture description has traditionally been regarded as advantageous compared to other less structured discourse production tasks, such as conversational or procedural tasks (Ulatowska et al., 1988). Indeed, this method offers a sturdy and straightforward way to gauge discourse production, commonly utilized for analyzing discourse in AD. It enables comprehensive scrutiny of information content and facilitates the comparison of semantic elements concerning individuals, actions, and items. Notably, this approach doesn't demand specialized terminology and can be carried out using vocabulary typically learned early in life (Hirsh and Ellis, 1994). Among other more cognitive complex tasks, telling a story is popular. This is often done by retelling a well-known tale such as Pinocchio, or Cinderella and integrating the characters and events within a temporal framework, or alternatively, creating a story from story picture prompts (Toledo et al., 2017). Reading (with or without recalling), narrating personal life experiences or dreams etc., represent other commonly used tasks. Petti et al. (2020) and de la Fuente Garcia et al. (2020) offered a detailed review of studies on the matter.

Finally, in several cases, patient-doctor conversations have been recorded during clinical assessment or cognitive examination. For instance, in Mirheidari et al. (2019), neurologists were given instructions to follow a predefined set of questions specifically designed to uncover common signs of impairments during the conversation. This included the use of closed questions that required long-term memory recall, compound questions, open-ended questions and questions related to memory. Weiner and Schultz (2016) analyzed semi-standardized biographical interviews using the ILSE corpus. Espinoza-Cuadros et al. (2014) recorded speech from both clinicians and patients during structured interviews used in the administration of the Spanish version of the MMSE. Luz et al. (2018) analyzed

natural conversations extracted from the Carolina Conversations Collection (Pope and Davis, 2011), involving 21 ILWD and 17 individuals with no AD or neuropsychological diseases (interviewers were gerontologists and language students or researchers).

2.2. Psychometric-clinical indices

There are a number of methods and scales to evaluate the severity of cognitive impairments, dementia and AD, each with their own merits according to need, such as research or clinical practice, and the aspect being assessed. Measures for overall outcomes include, for example, the Global Deterioration Scale (GDS, Reisberg et al., 1982) and the Clinical Global Impression of Change (CGI-C, Guy, 1976). To assess functional ability and quality of life, there is the Progressive Deterioration Scale (PDS, DeJong et al., 1989) or the Disability Assessment for Dementia (DAD, McIntyre, 1994; Gélinas et al., 1999). Cognitive ability is evaluated using the Alzheimer's Disease Assessment Scale - cognitive subscale (ADAS-cog, Rosen et al., 1984), the Severe Impairment Battery (SIB, Panisset et al., 1994), the Montreal Cognitive Assessment (MoCA, Nasreddine et al., 2005) or the Mini-Mental State Examination (MMSE, Folstein et al., 1975).

The MMSE is probably still the most used index today in clinical practice. It is a rapid administration test (10–15 min), composed of 30 items exploring temporal-spatial orientation, short-term memory, attention and calculation, ability to recall memories, comprehension and language, nominal ability, and the ability to write legibly and to copy minimally complex designs.

On the whole, the literature suggests that the MMSE exhibits internal consistency short-term test-retest reliability in ILWD and long-term reliability in those with intact cognitive functioning. Furthermore, the MMSE has been found to be sensitive to the severity of dementia in AD patients. The total score proves valuable in documenting cognitive changes over time, with AD patients typically experiencing an annual decline of 3 points on the MMSE (Bernard and Goldman, 2010). However, it is important to note that the MMSE is not used as the sole criterion for diagnosing dementia (Creavin et al., 2016; Arevalo-Rodriguez et al., 2021), also because of the potential influence of non-neurological factors on low scores (e.g., low education, language difficulties, visual or auditory impairments). Other problematic aspects and limitations of MMSE are discussed in Mitchell (2009), Tsoi et al. (2015), De Roeck et al. (2019).

Apart from its routine use in clinical practice, the MMSE is also employed as an exclusion or inclusion criterion in clinical trials (screening for cognitive impairment) and is incorporated into research study neuropsychological test batteries. In Italy, the MMSE is required for the accreditation of Nursing Homes to its National Health Service. The MMSE is usually administered to patients within 30 days of admission and subsequently every 6 months in accordance with the Individual Care Plan, except for unexpected developments that require new evaluation. Each Italian Region encompasses the national law into its own Regional Council resolutions (e.g., in Lombardy, DGR 7435/2001 and DGR 1765/2014).

The MMSE score is also adopted to define a scale of severity of AD, as for example the scale reported in note 85 by AIFA (the Italian Medicines Agency), which authorizes the prescription of AChE inhibitors (donepezil, rivastigmine and galantamine) for mild and moderate AD, and memantine for moderate and severe AD. In the same note, the following staging based on MMSE is reported: AD mild (MMSE 21-26), moderate (MMSE 10-20), moderately severe (MMSE 10-14) and severe (MMSE <10). The same scale is also reported by the National Institute for Health and Care Excellence (NICE), England.

2.3. Analysis by means of digital linguistic biomarkers

Correlating the linguistic production of speakers with cognitive impairments to clinical indexes, for example MMSE, may be pursued by means of various methods.

In various studies, linguistic features have been utilized, revealing correlation values that vary significantly with the type of task or language, playing a fundamental role. On the other hand, it is frequent in recent literature to call the linguistic characteristics useful for the identification of diseases as "Digital Linguistic Biomarkers". A comprehensive overview of the features and their hypothesized discrimination power in dementia research was presented in de la Fuente Garcia et al. (2020: 1552) and Petti et al. (2020). In particular, the ADReSS Challenge at INTERSPEECH 2020 (Luz et al., 2020) defined two cognitive assessment tasks: a standard Alzheimer's speech classification task and a neuropsychological score regression task. In the latter, participants had to create models to predict MMSE scores, using audio recording (and annotated transcriptions) elicited from participants using the Cookie Theft picture (Luz et al., 2021a, 2021b). A combination of audio and linguistic features extracted directly from audio recordings yielded a baseline MMSE prediction root mean squared error (RMSE) of 5.28 (Luz et al., 2020). A number of participants reported and discussed the results of the MMSE regression task. The best performing models reached the following scores: Koo et al. (2020), RMSE 3.75 (using a combination of audio and textual features); Balagopalan et al. (2020), RMSE 4.56; Haulcy and Glass (2021) RMSE 4.56; Meghanani et al. (2021), RMSE 4.28; Millington and Luz (2021) RMSE 5.46; Shah et al. (2021), RMSE 5.62.

In Yeung et al. (2021), a set of features provided by NLP and automated speech analysis was extracted from 30 speech samples with AD (MMSE: 15-20), MCI (MMSE: 23-26) and healthy controls (HCs, MMSE: 27-30), with a minimum education level of 12 years. The same speech samples were rated by 5 clinicians with respect to (1) word-finding difficulty, (2) incoherence, (3) preservation and (4) errors in speech, on a Likert scale (0=not present or normal finding; 1=mild; 2=moderate; 3=severe). The Authors found statistically significant correlations between a number of features and the clinical rates.

Along similar lines, Bueno-Cayo et al. (2022) evaluated the correlation between MMSE scores and linguistic features in a group of 33 participants (20 HCs and 13 individuals with MCI, aged between 60 and 95; up to a maximum education level of 12 years). They found that the lexical density (estimated as the ratio between the number of semantic content words per 100 words in a sentence) was positively correlated with MMSE scores (Pearson's correlation $= 0.488$, $p < 0.01$). MMSE scores were also negatively correlated with age. No statistically significant correlation ($\alpha = 0.05$) was found between MMSE and speech length, word frequency, and the number of

time-, place-, or action-related tokens. Lexical density and speech length were found as significant linear predictors in a multivariate linear fit of MMSE. Speech was elicited by a single open question: "what are your plans for today and what are your plans for tomorrow", with no constraint on the response time.

However, Kavé and Dassa (2017) found that MMSE scores were negatively correlated (Pearson) to the total number of words (-0.355, $p < 0.5$) and to the mean word frequency (-0,339, $p < 0.5$, greater dementia severity was associated with production of more frequent words); and positively correlated to the (types to token ratio) TTR (0.572, $p < 0.01$) in 35 individuals with AD (23 female, 12 male; age: 65–91; MMSE: 3–25; education: 8–16; Cookie Theft picture description task), while no statistically significant correlations were found to content-word ratio, pronouns ratio, percentages of verbs, prepositions, or to subordination markers.

Ostrand and Gunstad (2021) found that in the picture description task the use of definite articles, determiners, and nouns as percentage of total word count were each positively correlated with the MMSE scores (3MS, Teng and Chui, 1987), and that the average lexical frequency was negatively correlated to 3MS. Crucially, in an expository speech task by the same individuals, only the latter metric was found to be correlated to the 3MS.

These examples demonstrate how complex the scenario is. There are numerous DLBs which may be more or less correlated with clinical indexes, and these will all depend on a number of other factors, among which the type of task and linguistic focus play a crucial role. For example, in the work by Ostrand and Gunstad (2021), only lexical density correlated with 3MS in both the tasks discussed. Moreover, it should be noted that to date no strong correlation between cognitive decline and LCIs has been observed, given the scarcity of studies correlating language and MMSE. A few notable exceptions may be mentioned, mostly consisting of works focused on subjects suffering from AD (see, e.g., Fritsch et al., 2019) rather than general cognitively impaired subjects.

Most studies, aiming at identifying AD or MCI, have proposed categorization tasks and statistical comparisons of DLBs between a group of unhealthy people (either AD or MCI) and a group of matched HCs, as in Beltrami et al. (2018). Even when datasets include three or more diverse cohorts, most studies tend to only perform pairwise comparisons, typically contrasting dementia with healthy aging (Gagliardi, 2023). However, from a clinical application standpoint, this approach may not be useful for real-life situations. In fact, the prevailing focus on pairwise comparisons fails to fully grasp the complexity involved in assessing cognitive frailty in geriatric settings (Panza et al., 2015).

A number of studies have (either exclusively or in addition) compared "grades" of cognitive impairments, which is somewhat closer to the present work. For instance, König et al. (2015) considered AD, MCI and HCs (vocal tasks were: counting backwards, sentence repeating, picture description, verbal fluency) However, they only found significant differences ($p < 0.05$) between AD and MCI in the neuropsychiatric inventory (Cumming et al., 1994) Just a limited number of research papers are dedicated to the detailed classification of dementia, and these mainly focus on the subtyping of frontotemporal degeneration (e.g. Fraser et al. 2014; Garrard et al. 2014; Nevler et al. 2019; Themistocleous et al. 2018, 2021; Cho et al. 2020).

On classification techniques, regarding studies on the Italian language among the many contributions, L. Calzà et al. (2021) and Gagliardi and Tamburini (2021, 2022) explored various Machine Learning algorithms (such as Support Vector Machines, Random Forests, and Decision Trees) to automatically differentiate HC from MCI, based on features extracted from audio and (manual and/or automatically) transcribed speech, such as lexical, syntactic, semantic, and readability indexes.

2.4. Sentiment and emotion analysis

The progression of dementia decreases cognitive abilities and functional skills, preventing individuals from engaging in their usual daily activities. Moreover, subjects affected by AD may experience behavioral and social skill deterioration, leading to possible conflicts with others and ultimately to social isolation which further affects their emotional well-being (Logsdon et al., 1999).

Consequently, emotion analysis in AD speech can be of great help in monitoring the degree of disorder and in distinguishing ILWD from HCs. In the latter case, for example, López-de-Ipiña et al. (2015) achieved a classification error of less than 5 %, using Emotional Temperature and varied use of the fractal dimension (such as Higuchi, 1988).

Traditional approaches in speech emotion recognition consider linguistic and paralinguistic features extracted from the speech signal (Altun and Polat, 2009). On the other hand, modern LMs can also be used to analyze sentiment and emotional content of transcribed speech. In a recent study, Liu et al. (2023) has proposed an innovative attention mechanism to detect deep semantic information within words and sentences, to clarify the meaning of the transcribed discourse. Bianchi et al. (2021) presented FEEL-IT, a benchmark corpus of Italian Twitter posts annotated with four basic emotions: anger, fear, joy, and sadness (Ekman, 1992). The Authors also used the Italian BERT model UmBERTo[1] trained on Commoncrawl ITA, then fine-tuned to classify emotion using the FEEL-IT corpus. Given an input sentence, the model can provide an output probability for the four emotions and an output label for the emotion with the highest probability. Additionally, collapsing anger, fear, and sadness into the "negative sentiment" category, and joy as "positive sentiment", the model can perform sentiment analysis. The model achieved F1 = 0.80 on the 2-sentiments classification task, and F1 = 0.57 on the 4-emotions classification task.

2.5. Analysis of perplexity

One successful application of LMs in discriminating classes of subjects based on their verbal output relies on the adoption of

[1] Common Crawl is an open repository consisting of web crawl data that has been collected since 2008 and can be freely accessed and analyzed. For UmBERTo, see https://github.com/musixmatchresearch/umberto.

perplexity (PPL), which measures how unlikely a given text sequence is (Goldberg, 2017), with respect to a given LM (more details in Section 4.2.3). The basic assumption is that if we train an LM based on the language produced by healthy subjects, and if we then use that LM to calculate the perplexity values of two previously unseen text sequences, one emitted by a healthy subject and one emitted by a cognitively impaired subject, we would expect to observe a higher perplexity score featuring the text authored by the impaired subject. This assumption allows for variances from the expected pattern, and it has been proven successful (Colla et al., 2022) in specific experimental conditions, such as the categorization of unhealthy subjects and controls based on the descriptions for the Cookie-theft picture task in the Pitt-Corpus.

In their pioneering approach, Solorio and Liu (2008) trained two LMs (based on N-grams) with data from patients and HCs respectively, then used them to categorize a speech sample of a test subject by computing the perplexity score for both models and choosing the corresponding class that gave the lowest value. Fritsch et al. (2019) trained two neural LMs, using the transcriptions from the Pitt Corpus for HCs and ILWD, to categorize speech production in the two groups. The authors achieved 85.6 % accuracy on 499 transcriptions and, perhaps more importantly for the present work, showed that perplexity can also be exploited as a predictor for patient MMSE scores. In a very similar classification experiment, Cohen and Pakhomov (2020) achieved 0.872 accuracy at equal error rate. Furthermore, such experiments showed that as the disease advances, the perplexity of neural LMs grows consistently; in particular, the experimentation provided evidence about the correlation among perplexity scores and lexical frequency. In fact, higher-frequency and less specific words occurred more frequently in the language of ILWD than in that of the HCs.

The work in Colla et al. (2022) is twofold. First, it explores the reliability of the perplexity metrics, by testing whether perplexity is consistent enough to analyze the language of individual subjects, and still sensitive enough to capture language and register changes made by a single speaker in different communicative situations, such as that of an interview rather than a political rally.[2] Secondly, in this work a new method was proposed to refine the decision rule to categorize ILWD and HCs[3]: The method successfully ranked the models, with the best achieving full accuracy and F-score when using the Cookie-theft picture task from the Pitt-Corpus.

3. Goals of the investigation

As mentioned in Section 1, the main goal of this work was to test the effectiveness of a number of automatic analyzes using transcribed conversations of individuals under conditions of high ecological validity, and at different stages of cognitive impairments as denoted by their MMSE score.

In order to achieve such a goal, various sub-goals had to be reached:

1. The analysis of different automatic analysis methods to characterize the speech of subjects affected by dementia, considering:
   a. various types of information, from lexical and morphosyntactic choices to variations in attitude, including sentiment and emotion;
   b. classical statistical methods, machine learning approaches and more recent deep learning based language models.
2. The distinction of different stages of impairment by identifying the way transcribed speech characteristics:
   a. correlate with MMSE scores;
   b. can be classified in terms of MMSE clusters and levels of severity, as identified in the literature (with specific reference to the AIFA classification);
3. The in-depth analysis of naturalistic data, as those offered by the Anchise Corpus.

4. Materials and methods

4.1. Dataset description: the Anchise 2022 Corpus

The corpus has been collected since 2007 by the Anchise Group,[4] an association of experts for the research, training and care of ILWD. It is based on the adoption of the *ApproccioCapacitante*®, or the 'Enabling Approach', developed by one of the Authors and his collaborators since 2004 (Vigorelli, 2004) and subsequently refined (Vigorelli, 2010, 2011; Lanzoni et al., 2018; Vigorelli, 2021, and 2024). This approach is a 'non-pharmacological therapy' of dementia designed to improve communication in the presence of evident memory and language disorder. Its ultimate goal is to promote a sufficiently happy coexistence between ILWD and others. To achieve this goal, ILWD engage in conversation with a wide variety of operators (health care workers, educators, nurses, speech therapists, doctors, psychologists) who have previously undergone training in the *ApproccioCapacitante*.

Following the *ApproccioCapacitante* protocol, after the greetings and the collection of informed consent, the operator encourages the patient to speak, with suggested openings such as "Could you tell me something about how you are going to spend your day?", "What do you do during the day?", "We have some time to get to know each other better", or "Can we talk together so that I can get to know you better?". Then the operator should mainly listen, waiting for the ILWD to speak. Only when there is definite pause should the operator take the floor and continue the conversation following on from what the other has said. In general, the operator does not lead

[2] The set of speech transcripts collected for such experiments are described in the work by Colla et al. (2023).

[3] Namely, perplexity scores averaged over the two classes were combined with deviations, based on the 3σ rule, a popular heuristic in empirical sciences (Helms, 2009).

[4] www.gruppoanchise.it

the conversation but rather follows the patient and accepts their understanding of the world. The operator is categorically not involved in making a diagnosis or even in gathering information, but rather participates in the conversation in a way that promotes the well-being of both speakers during the conversation itself. The main rules to follow when the ILWD take the floor are[5]: "Don't ask questions", "Don't interrupt", "Don't correct", "Echo", "Return to the main conversation topic". Other instructions are "Listen carefully even when speaking is pathological and incomprehensible", "Respect any sluggishness and pauses", "Recognize and mirror emotions", "Answer questions", "Also use gesture and tone of voice to communicate" (Vigorelli, 2004, 2011). Though there are no rules regarding the duration, the conversation usually lasts 5–10 min, ending when signs of tiredness or intolerance appear in the patient. Crucially, these instructions are given to the operators in a training context, where the protocol is strictly followed, and may be flexibly adopted in the field.

Since 2007, health professionals from the Anchise Group have recorded, manually transcribed, and annotated conversations involving elderly people with cognitive impairment mostly residing in Italian Nursing Homes. All the participants are Italian speakers with evident cognitive deficits (MMSE score range = 0–28), though with no psychological or behavioral disturbances that might hinder conversation, such as drowsiness, confusion, a markedly oppositional or aggressive attitude, evident psychomotor agitation, severe dysarthria or severe hypoacusia.

All the conversations carried out with the operators participating in the Anchise Group have been manually transcribed according to internal transcription conventions and included in the corpus, and importantly no selection was carried out based on adherence to the expected style. The material is strictly for training use with Nursing Home operators only, while the original audio recordings have been deleted. The corpus is being continuously updated by the Anchise Group. A previous version of the corpus, called 'Corpus Anchise 320' (Bolioli et al., 2020; Benvenuti et al., 2021), was released in 2020: it contained 320 conversations, while the Anchise 2022 Corpus contains 417 conversations.

The significance of a study based on the Anchise Corpus stems from a number of factors: it has been collected in a naturalistic environment; it includes dialogical speech; it is composed by a high number of recorded conversations, by as many ILWD; it is a corpus in Italian, and it offers a cross-section from North to South of the situation in Italian Nursing Homes. The corpus includes both large and small homes, and patients of all types, with mild, moderate and severe cognitive deficits, and from both sexes. The classification of patients is based only on the MMSE score. However, unlike prospective, controlled studies performed under standardized conditions, this corpus was not created for research purposes, but rather to be used in personnel training. As such, the corpus offers a substantial source of raw data on language produced by subjects affected by dementia.

4.1.1. *Descriptive statistics of the corpus*

Each dialogue or conversation is arranged into speech turns: a new turn starts each time the speaker changes and lasts until there is a significant interruption. We assume that each conversation involves a different patient. The two conversations out of the 417 carried out in other languages (Spanish and English) were discarded. Not all the conversations were accompanied with all the additional information concerning the patient, such as sex, age and related MMSE score.

We selected the transcripts of patients with an MMSE of up to 26 (see Section 4.2.1), and only included those with details of age, to have more control of the variables, resulting in a total of 216 conversations.

Table 1 illustrates some preliminary insights regarding the original and the filtered datasets. The patients interviewed average around age 84 (with a similar spread in both datasets), while the corpus in both datasets is dominated by women (almost 80 % of our speakers, see Fig. 1). Average number of turns is 65, almost equally split between Patient and Operator, with an average of 14+ tokens/turn. This figure significantly differs if we consider Patient turns (longer, around 20 tokens/turn) and Operator turns (shorter, about 9 tokens/turn). This figure is complemented by a high standard deviation, which is in the same order if not greater than the average itself. Conversations in the filtered corpus are slightly longer than those in the rest of the Corpus (843 vs. 831 tokens, complemented by a high standard deviation, thus suggesting that consistent variation in length may be found), with Patient conversations almost twice as long as Operator. As regard the relation between MMSE and AGE (see Fig. 1) their joint distribution is almost Gaussian (Henze-Zirkler test = 0.86, p-value = 0.15), while MMSE is (weakly) negatively correlated to AGE (β = -0.11, 95 CI = [-0.21, -0,00,426], t(217) = -2.05, p = 0.041).

4.2. *Methods*

After having decided on the clustering of MMSE values, two broad methods were adopted to characterize the transcribed conversations and the MMSE score: those employing language features and those relying on Language Models.

4.2.1. *MMSE clustering and evaluation measures*

We used the MMSE scale as recommended in note 85 by AIFA and NICE. For sake of simplicity, the term 'AIFA85' will be used hereafter to indicate the stages of cognitive impairments as follows:

- SEVERE: MMSE 0–9, 60 individuals.

[5] Our translation.

Table 1
Summary description of the Anchise 2022 Corpus composition with respect to subject type (patients, operators), patient sex, speech turns and availability of mini-mental state examination score and age data. Where appropriate, the values are expressed in the form: mean (standard deviation).

	Unfiltered Corpus	Filtered Corpus
Number of conversations	417	216
Average patient age	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	12.69 (5.60)	12.92 (5.46)
Average number of turns per conversation	65.3	64.90
Average number of patient turns per conversation	32.3	32.04
Average number of operator turns per conversation	33	32.86
Average turn length	14.25 (10.12)	14.83 (12.05)
Average turn length Patient	19.60 (19.43)	20.31 (23.31)
Average turn length Operator	8.85 (3.58)	9.24 (3.69)
Average conversation length	831.68 (632.60)	843.50 (535.21)
Average conversation length - Patient	543.76 (456.02)	545.82 (395.64)
Average conversation length - Operator	287.92 (236.32)	297.67 (214.42)

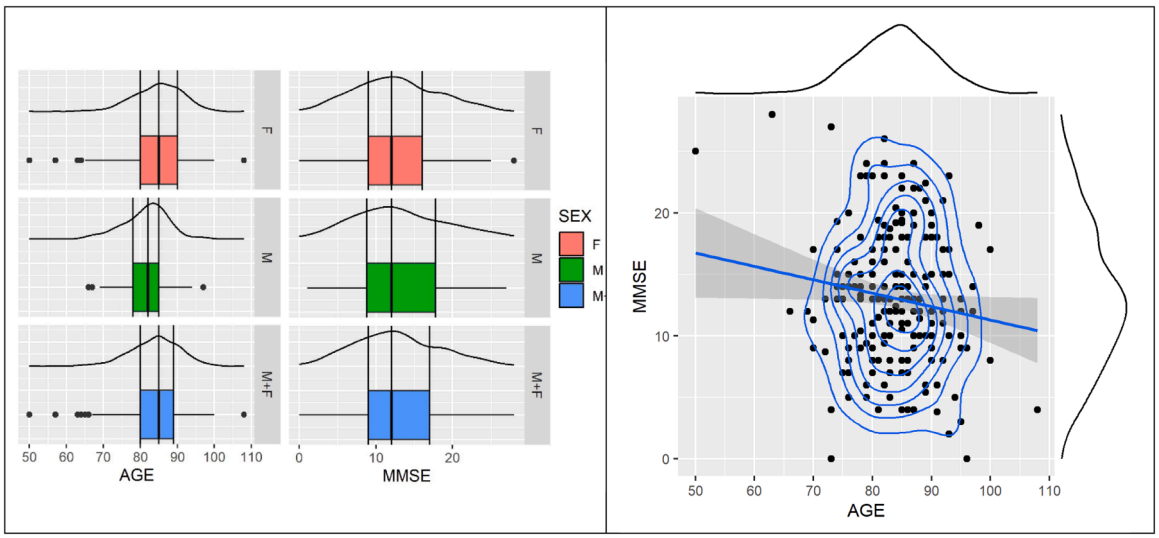


Fig. 1. Unfiltered corpus distribution of the number of conversations with respect to sex, age, and MMSE scores. On the left: boxplots with kernel density estimation for AGE and MMSE by sex: female, upper plots), male (central plots) and both (lower plots). On the right: scatterplot (dots), bivariate contour lines, linear trend, and marginal kernel density estimations, for age and MMSE score without sex distinction.

- MODERATE: MMSE 10–20, 137 individuals.
- MILD: MMSE 21–26, 19 individuals.
- NORMAL: MMSE 27–30, 2 individuals.

Experimentation considered only SEVERE, MODERATE and MILD stages, since the category of MMSE 27–30 contained only 2 individuals.

Finding correlations with MMSE scores amounts to finding a set of evaluation tools that can predict the MMSE most probably associated with the patient involved in that conversation. Below we compute the Pearson and Spearman correlation coefficients (Benesty et al., 2009).

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. Also, 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.
- an 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. The same applies to macro Recall and macro F1.

4.2.2. Linguistic biomarkers

Using NLP tools provided by Stanza software (Qi et al., 2020), a set of morphosyntactic data and dependency relations among the parts-of-speech (POS) were extracted from each turn in each patient conversation (see Table A-1, A-2 and A-3 and Fig. A-1 in the Appendix). Universal POS Tags (UPOS) and language-specific POS Tags (XPOS) for Italian have been used in this work.

A number of DLBs (at lexical, syntactic, and semantic levels) were then analyzed for each conversation (see Table A-4 for more details). Consequently, each conversation was allotted its own vector of DLBs.

4.2.2.1. Correlation and classification of DLBs and MMSE scores. Each feature was correlated to MMSE scores with linear fit, Pearsons and Spearman.

The features were then used for statistical comparisons between clusters of MMSE scores based on the AIFA85 stages (Section 4.2.1). Because of the unequal and relatively small sizes of the sample for some of the stages, the non-parametric Kruskal-Wallis as omnibus test and the two-sample Kolmogorov-Smirnov (KS) test was chosen, which estimates a distance between the empirical distributions of the single features in both classes. An α=0.05 significance level was set as the threshold to assess statistical significance.

Additionally, a classification task, predicting the class of a conversation based on its DLBs was performed using the tool illustrated in Fig. 2. The tool is a three-class classifier, composed of three binary classifiers, each one devoted to predicting two classes. The first classifier signals that the input (vector of linguistic features of that) session is SEVERE or MODERATE, the second classifier signals SEVERE or MILD, while the third signals MODERATE or MILD. The final prediction for an input conversation is chosen as the most "voted" output class.

Each binary classifier is trained separately from the others, using only the conversations belonging to the two signaled classes, and

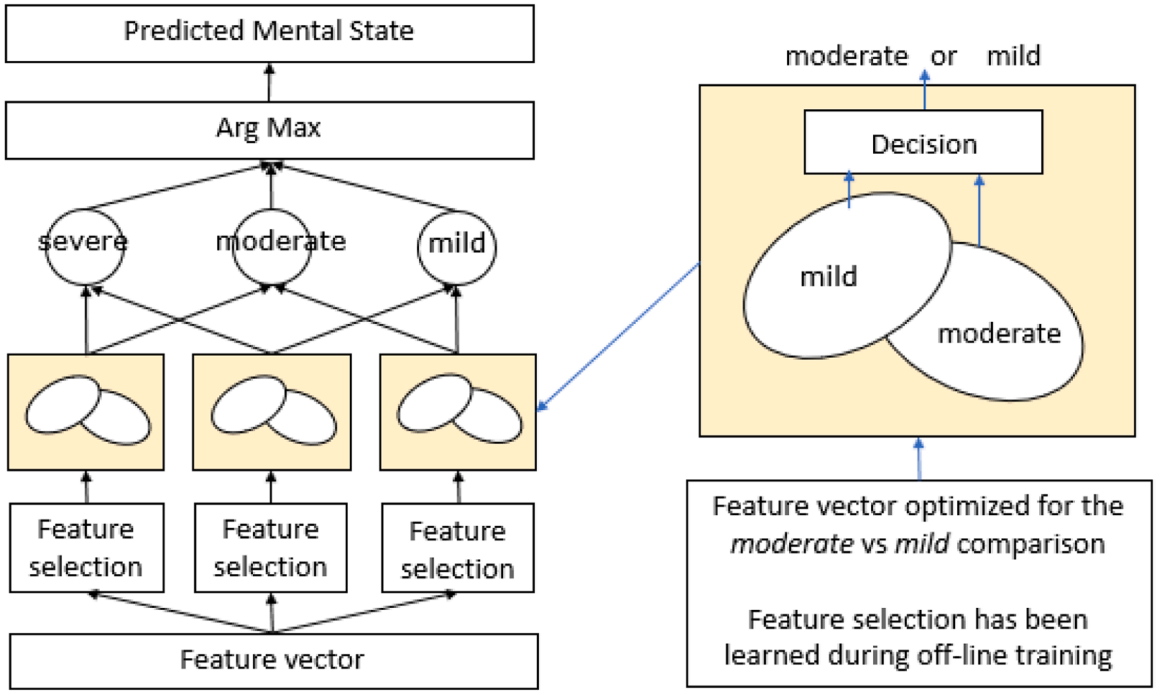


Fig. 2. Scheme of the multiclass classifier used for the language features classification experiment. On the left: the multiclass classifier, composed of three binary classifiers. On the right: the single binary classifier. Feature vector: the vector of digital linguistic biomarkers for each conversation. The training of each binary classifier consists of learning which features need to be selected (and which to discard) to improve classification accuracy, as well as the type and final parameters of the classifier. In the testing phase, each conversation's feature vector is sent to three binary classifiers. Once each of the classifiers has made its own feature selection, it also decides on a class. The most voted class will be the output of the multiclass classifier (predicted mental state).

using a leave-one-out cross-validation (LOOCV) approach. In other words, each conversation is extracted in turn from the data set and used for prediction, while the remaining conversations are used as a training data set.

Two types of binary classifiers have been tested. The first assumes that the statistical distribution of the DLBs for each of the two classes is Multivariate Normal (MVN), meaning that the probability distribution is completely determined by the vector of means and the matrix of variances and covariances. For each of the two classes, the vector and matrix are estimated during the training phase by using the feature dataset for each class. At the test stage, given the input vector of features, the binary classifier outputs the class with the higher probabilistic likelihood. The second type of binary classifier, is based on the well-known Linear Discriminant Analysis (LDA), as implemented by the Matlab® *fitcdiscr* function.

During training, feature selection algorithms, custom implementations of Forward Feature Selection (FFsel) and Forward-Backward Feature Selection (FBsel) were also employed, to maximize the performance metrics, and to highlight the set of features most useful for the classification.

The Forward Feature Selection procedure starts with an empty set of selected features, and then populates this set progressively, considering the available features one by one, from first to last, each only once. A new feature is added to the set of selected features only if this new set provides a better classification performance than that obtained in the previous step (i.e., without the new feature). In this phase, the macro average accuracy is considered as the performance metric, the arithmetic average of the true positive rates. To evaluate classification performance, the main component of the binary classifier (either MVN or LDA) is trained with the current set of selected features and tested by using the LOOCV approach.

The Forward-Backward Feature Selection is similar to the FFsel, with the difference that every time a new feature is added to the set of selected features the algorithm checks whether the performance can further increase through the elimination of one or more of the previously selected features. If this happens then the identified features are eliminated from the set of selected features.

It should be pointed out, however, that neither of the two feature selection methods guarantees any absolute best set of features (other combinations giving better or equal performances may still exist).

In summary, the training phase for each binary classifier consists of an initial feature selection phase, and finally, once the "best" features have been learned, the classifier is trained one last time with these selected features.

In the testing phase, the full vector of test features (corresponding to a new conversation) is firstly sent to each binary classifier, where it is subjected to the respective feature selection (already learned in the training phase). The resulting feature vector is sent to the classifier core which returns the most likely class. The output of the multiclass classifier is finally composed of the most voted class from the three binary classifiers.

4.2.3. Language models and perplexity

An LM or language model (Manning and Schutze, 1999) is a statistical inference tool that can estimate the probability of a word sequence $W = \{w_1, \ldots, w_k\}$, for any possible sequence of words W (Goldberg, 2017). Such probability can be computed as

$$P(W_{1,n}) = \prod_{i=1}^{k} p(w_i | w_1, \ldots, w_{i-1}),$$

which is customarily approximated as

$$P(W_{1,n}) \approx \prod_{i=1} p(w_i | w_{i-N+1}, w_{i-N+2}, \ldots, w_{i-1}),$$

where the entire sequence W is predicted based on blocks of exactly N words (i.e. N-grams).

The probabilities assigned by LMs are the outcome of a learning step, in which the model is exposed to a given set (and thus to a particular kind) of textual data. The goal of the training phase is to train the model to predict word sequences similar to those encountered during training. This feature constitutes the main trait of models such as BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) and GPT-2, Generative Pretrained Transformer-2 (Radford et al., 2019), which have both been successfully adopted in many mainstream tasks, such as Named Entity Recognition, Textual Entailment, Coreference Resolution, Paraphrase, Sentence Similarity, Natural Language Inference and Question Answering (Wang et al., 2018, 2019; see also the literature review by Colla et al., 2020).

In this setting the most influential approaches employed to analyze the language of cognitively impaired subjects concern the adoption of the perplexity metrics. Perplexity is a positive score that expresses how unlikely it is for a model to generate a given sequence, i.e., how perplexed the model is in emitting that sequence: low values (corresponding to high probability values) indicate that the model is able to predict that sequence. Formally, the perplexity (*PPL*) of a language model LM with respect to a word sequence $W = \{w_1, \ldots, w_k\}$ is computed through the following equation:

$$PPL(LM, W) = exp\left\{ -\frac{1}{k} \sum_{i=1}^{k} log\, LM(w_i | w_{1:i-1}) \right\}.$$

Perplexity has been widely used to compare text sequences, and to discriminate between sequences produced by healthy subjects and sequences produced by people suffering from language-related disturbances (Fritsch et al., 2019; Cohen and Pakhomov, 2020; Colla et al., 2022).

4.2.3.1. Correlation between perplexity scores and MMSE scores. In one of our experiments, we explored the correlation between the perplexity measure and the cognitive impairment level assessed by the MMSE scores.

We first pre-processed each transcript by concatenating all the sentences produced by the patient only, $t^p = \{w_1\ w_2 \ldots w_n\}$, for the patient p, where w_i are the words spoken. We then provided each t^p with a perplexity score, calculated by the language model LM. To do this, we computed a list of perplexity scores $PPLI(LM,\ t^p) = [s_1,\ \ldots,\ s_n]$, where s_i was calculated for each w_i, in an incremental fashion, as illustrated in Eq. (1):

$$s_i = PPL(LM,\ w_i \mid w_{<i}\ )\ \forall w_i\ \in\ t^p, \tag{1}$$

where $PPL(LM,\ w_i \mid w_{<i}\ )$ indicates the perplexity score calculated by the language model LM for the word w_i given the preceding context $w_{<i}$.

At this point, the score for the patient p may be obtained by averaging the perplexity s_i in $PPLI(LM,\ t^p)$, as follows:

$$PPL^p = \frac{1}{n}\sum\nolimits_{i=1}^{n} s_i\ \forall s_i\ \in PPLI(LM,\ t^p).$$

However, because of the dialogic nature of the transcripts, the opening conversation turns contain greetings. Even though greetings are appropriate for a conversation, these might be misleading for a language model unfamiliar with this genre, and consequently might create noise in the computation of perplexity. Hence, we investigated the effect of removing a small initial X% of turns. In other words, the perplexity of a transcript was calculated using Eq. (2):

$$PPL_X^p = \frac{1}{n\ -\ \left(\frac{n\ -\ X}{100}\right)} \sum_{i\ =\ \frac{n\ -\ X}{100}}^{n} s_i\ \forall s_i\ \in\ PPLI(LM,\ t^p) \tag{2}$$

as the average of the scores in list $PPLI(LM,\ t^p)$ except for those first X% of words that were dropped. The (negative) correlation between MMSE and PPL scores is reported in the following, in Section 5.1.2, Fig. 2, and shows that by filtering out the initial stages of the conversation produces increased correlation.

To compute the perplexity scores, we employed an LM for the Italian texts, GePpeTto (De Mattei et al., 2020), which is based on GPT-2 architecture (Radford et al., 2019). Given the dialogic nature of the texts within the corpus, we additionally investigated the effect of coaching the GePpeTto model, already pre-trained in Italian, on texts of a similar nature, using the ParlaTO dataset (Mauri et al., 2019). This dataset contains about 50 h of dialogues involving over hundred speakers of various ages, collected in Italy between 2018 and 2020. To provide as much consistency as possible with the Anchise Corpus, we only selected the transcripts of conversations with people aged over sixty. The model was then trained on 384,956 tokens for 10 epochs.

Finally, to assess the relation between the perplexity and the MMSE scores we employed the Pearson correlation index.

4.2.3.2. Classification through language models and perplexity. An array of experiments was devised to explore to what extent LMs can be used to predict the degree of severity of cognitively impaired subjects. Namely, we experimented with the multilingual version of the BERT uncased base model trained on a subset of the ParlaTO corpus containing dialogues involving speakers aged over 60 (Mauri et al., 2019). The model has been trained on 384,956 tokens for 10 epochs with a batch size of 8 instances, which took about an hour. The multilabel categorization (classes: MILD, MODERATE, SEVERE) on the Anchise Corpus was complemented by an experiment employing a balanced variant of the Corpus.

Furthermore, we performed two more experiments casting the multilabel categorization task to a set of binary categorizations, where three binary classifiers (Mild-Moderate, Moderate-Severe, Severe-Mild) were acquired and tested; in this setting we compared the results obtained by employing standard BERT-based categorization with a classifier employing perplexity scores. A layer consisting of two units with Gaussian Error Linear Unit (GELU) activation was placed on top of the BERT model to build the classifier. Each model was fine-tuned for ten epochs, with a batch size of eight instances; a ten-fold cross validation setting was employed. The AdamW optimizer was chosen, equipped with cross-entropy as the loss function. Learning rate was set to 1e-03, and the eps (the AdamW epsilon) was set to 1e-06.

4.2.4. Sentiment and emotion analysis

To build a picture of sentiment and emotions using the Anchise Corpus, the FEEL-IT model for Italian was used (Section 2.4). By providing a conversation turn as input text string, the model calculates the probabilities of anger, fear, joy, and sadness. This calculation was repeated for all turns within a single conversation with the elderly patient to find and label the most probable emotion for each turn. By collapsing this information, we obtained a rate of occurrence for each emotion, as the number of occurrences of speech turns labeled with a given emotion divided by the total number of turns in that conversation. By repeating these steps for each conversation, a set of tuples was obtained, in which the MMSE value of each conversation can be considered as an independent variable, while the associated emotion rates can be considered as the dependent variables.

The results were then used for subsequent analyzes. In particular, the correlation (Pearson's and Spearman's) between the rate for each emotion and the MMSE score were explored (Section 5.1.3). Furthermore, the proportion of occurrences of each emotion in the SEVERE/MODERATE/MILD categories was statistically analyzed (Section 5.2.3).

As for the sentiment analysis, this was achieved through a second model using FEEL-IT, which merges anger, fear, and sadness into a single class of negative emotions, against joy, the only positive emotion (Bianchi et al., 2021). Starting from the procedure illustrated

above, the polarity of sentiment (positive or negative) was then initially calculated for each turn of speech. Unlike emotion analysis, for sentiment analysis we decided to give a sentiment label at conversation (i.e. individual) level. To be labeled positive or negative, there needed to be a majority of at least two. In cases where the majority was only one or there was an exact positive/negative balance the conversation was labeled as neutral, thus creating a new sentiment at conversation level. This requirement arises from the obvious observation that in the case of a conversation with an even number of turns an equal number of "positive" and "negative" turns can occur, while this is absolutely impossible in conversations with an odd number of turns. The choice made seemed to be a good compromise to handle these eventualities. Regardless of the sentiment label assigned to a conversation, within each conversation we also calculated the *pos_rate* as the number of positive turns divided by the number of turns in the same conversation.

Having characterized each conversation with sentiment and emotion features, possible connections between emotions and the MMSE in the Corpus could be studied (Sections 5.1.3 and 5.2.3).

5. Results

5.1. Correlations with single MMSE scores

5.1.1. Correlation between DLBs and MMSE

In general, all the correlations between MMSE and each single DLB were found to be relatively weak. Nevertheless, some were statistically significant ($p < 0.05$) and are shown in Table 2.

As for positive correlations, the count of hapax legomena (the words spoken just once in a conversation), the count of negative adverbs (BN) and verbs in the subjunctive mood are significantly correlated according to both Pearson and Spearman coefficients. The latter feature has the highest values of positive correlations (0.1902 for Pearson and 0.1910 for Spearman correlation coefficients). The frequency of verbs in the past tense, of modal verbs (VM) and the mean value of dependency distance are significantly and positively correlated according only to the Pearson's correlation coefficient. The noun (N) rate, the modal verb count (VM) and the reduced sentence count, marked by the number of participle and gerund verbal forms, are significantly and positively correlated only to Spearman.

As for negative correlations, the frequency of interjections (INTJ), finite verbs, present tense verbs and the rate of exclamative determiners (DE) were significantly correlated according to both the Pearson and Spearman coefficients. The latter feature has the highest values of negative correlations (-0.2069 for Pearson and -0.1438 for Spearman). The adjective rate (calculated with both ADJ and A POS tags) was significantly negatively correlated according to the Pearson coefficient only.

The rate of exclamative determiners (DE) is the feature that exhibits the best, albeit not strong, correlation with the MMSE scores. The following DEs have been found (together, they constitute only 0.1 % of the total number of words):

- 210 occurrences of *che* 'what', as *in che bel lavoro!* 'what a great job'
- just 12 occurrences of *quanto(a)* 'many', as in *guardi quanta gente!* 'look at all the people!'

Many other features showed no statistically significant correlation with MMSE. However, attention should be paid to speech style and to task related features. For instance, content density seems to be very sensitive to the language task. For instance, in Beltrami et al. (2016) content density was found statistically significant in the comparison between MCI and healthy controls in the picture description task, but not in the narrative task.

Table 2
Linear fit and corresponding statistical significance between each single linguistic feature and mini-mental state examination scores. Only statistically significant results are reported (* $p < 0.05$, ** $p < 0.01$). Abbreviations: beta, the coefficient of the linear fit; t, the value of the statistic, with 214 degrees of freedom; CI, confidence interval of the estimation of beta; RMSE, root-mean-square-error; R^2, R-squared, the fraction of the explained variance; r-Pearson and ρ-Spearman, correlation coefficients; reduced_sentences_count: the number of participle and gerund verbal forms; dep_dist.mean.avg: the mean value of dependency distance.

Features	Beta	t (214)	CI	RMSE	R^2	r	ρ
Hapax count	0.0132	2.2501	[0.0016, 0.0248]	5.40	0.0231	0.1520 *	0.1452 *
Adjective (ADJ) rate	-0.3909	-2.275	[-0.7295, -0.0522]	5.39	0,0236	-0.1537 *	-0.0693
Adjective (A) rate	-0.4013	-2.3449	[-0.7386, -0.064]	5.39	0,0251	-0,1583 *	-0.0794
Interjections (INTJ) rate	-0.1349	-2.4495	[-0.2434, -0.0263]	5.38	0,0273	-0.1651 *	-0.1605 *
Noun (N) rate	0.1798	1.6695	[-0.0325, 0.3922]	5,42	0,0129	0.1134 .	0.1595 *
Exclamative determiner (DE) rate	-2.3474	-3.0934	[-3.8432, -0.8516]	5.34	0,0428	-0.2069 **	-0.1438 *
Negative adverb (BN) count	0.0612	2.1061	[0.0039, 0.1184]	5.40	0.0203	0.1425 *	0.1577 *
Modal verb (VM) count	0.1332	1.8842	[-0.0061, 0.2725]	5.41	0.0163	0,1277 .	0.1492 *
Modal verb (VM) rate	1.0165	2.2552	[0.1281, 1.905]	5.40	0.0232	0.1524 *	0.1296 .
Finite verb rate	-0.077	-2.6926	[-0.1333, -0.0206]	5.37	0.0328	-0.1810 **	-0.1596 *
Subjunctive mood rate	0.2622	2.8343	[0.0798, 0.4445]	5.36	0.0362	0.1902 **	0.1910 **
Present tense rate	-0.0457	-2.2643	[-0.0855, -0.0059]	5.40	0.0234	-0.1530 *	-0.1416 *
Past tense verb rate	0.0650	2.0541	[0.0026, 0.1273]	5.41	0.0193	0.1391 *	0.1203 .
Reduced_sentences_count	0.0374	1.4477	[-0.0135, 0.0883]	5.43	0.0097	0.0985	0.1558 *
Dep_dist.mean.avg	1.8419	2.0996	[0.1127, 3.571]	5.40	0.0202	0.1421 *	0.1233

5.1.2. Correlation between perplexity scores and MMSE scores

The objective of this experiment was to investigate the correlation coefficients between subject MMSE score and conversation perplexity scores.

The results for this ParlaTO dataset are presented in Fig. 3. It compares perplexity levels of the unfiltered GePpeTto (blue line) with the filtered GePpeTto (red line) according to how much text has been filtered out (from 0 to 99 %) —i.e., from PPL_0^p to PPL_{99}^p for each patient p—.

We expected to find a negative correlation, hypothesizing reduced MMSE would result in a higher perplexity score, meaning that the model would be less confident, more perplexed, in predicting the language emitted by the patient suffering from a higher level of cognitive impairment.

This inverse correlation was observed both with the unfiltered and the filtered GePpeTto models on the ParlaTO dataset. At the beginning, filtering out the first 10 % of text, correlation reduces. Correlation then begins to climb, with a slight dip at around 30 %, to a maximum at 80 %. The effect of fine-tuning the model on a language close to the texts in the Anchise Corpus seems to improve the correlation, although not enough to ensure a clear positive result. However, this experiment does seem to confirm the fact that perplexity focusses on different aspects of language ability compared to the MMSE. In fact, while perplexity can only assess how close a linguistic sequence is to some reference language (such as that in the training data), the MMSE can assess a number of cognitive skills, including space and time orientation, attention and calculation, recall, language and praxis.

Perplexity has been experimentally proven as able to detect cases of AD using transcripts containing descriptions of the Cookie-theft picture (Colla et al., 2022). The experiments discussed here are rather different, in that they focus on naturalistic dialogues, where subjects talked freely to a healthcare operator, consequently making it harder to compare the conversations. We should also add that the corpus is made up of only 216 conversations, and the conversations themselves significantly differed in length (from less than 100 to above 2500 tokens). Overall, the 0.26 inverse correlation does not allow us to state that perplexity scores can be considered as a strong predictor of the MMSE score (or vice versa). What we can say is that perplexity scores are able to reveal more information on language, and that this detail can be considered as complementary to MMSE scores.

5.1.3. Correlation between sentiment and emotion analysis and MMSE scores

Observing the distribution of the positive labeled speech turns within a conversation (*pos_rate*) against MMSE values (Fig. A-2 (left)), we see a slight tendency towards a decrease in positive sentiment as the MMSE increases. Also, despite only explaining a small part of the variance (R^2=0.03), the trend is statistically significant (β = -5.01e-03, 95 % CI [-8.88e-03, -1.13e-03], t(214) = -2.55, p = 0.012; Std. beta = -0.17, 95 % CI [-0.30, -0.04]). The trend line moves from *pos_rate* values slightly above 50 % to values slightly below 50 %, which suggests that non-positive sentiment progressively increases with MMSE levels above 10–12.

The emotion analysis allows us to explore the results of the sentiment analysis in more depth. As depicted in Fig. A-2 (right), joy and sadness are the most frequent emotions, with a slight prevalence of "joyful" turns of speech over the "sad" ones. There are very few speech turns labeled "fear", while "anger" is close to 12 %.

As for the emotion rates at the conversation level (Fig. A-3 and Table A-5), we can note trends for anger, joy, and sadness, while fear is not statistically significant. As the MMSE increases, so does the anger rate (β = 8.4, Pearson c.c. = 0.15, p = 0.0176) as well as sadness (β = 5.7, Pearson c.c. = 0.16, p = 0.0161) while the joy rate decreases (β = -7.4, Pearson c.c. = -0.22, $p < 0.001$). This is in line with the sentiment analysis. The emotion of fear is mainly absent, and does not correlate with MMSE, while the other emotions show significant correlation both for Pearson and Spearman coefficients.

The results of the correlation analysis may be explained by hypothesizing a link between the emotions manifested and the degree of awareness of one's own state as in-patient and health. In this sense, as the MMSE increases, so does the degree of awareness, and consequently the rate of negative emotions (sadness and anger). On the contrary, as MMSE decreases, individuals tend to be less aware, and they can more easily recall positive emotion and better enjoy the benefits of conversing with enabling operators.

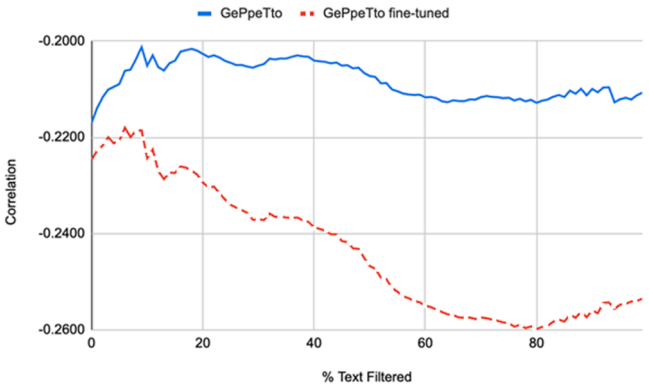


Fig. 3. Plot of the correlation scores between mini-mental state examination and perplexity calculated using Eq. (2). Correlation obtained by the GePpeTto model (blue solid line) and by the filtered GePpeTto model (red dashed line) on the ParlaTO dataset; Pearson correlation (vertical axis), percentage of transcript removed for perplexity (horizontal axis).

5.2. Classification

5.2.1. Classification via DLBs

A statistical analysis on the distribution of each language biomarker, independently of each other, across the AIFA85 stages, was performed using the Kruskal-Wallis (KW, as omnibus test) and the Kolmogorov-Smirnov (KS) tests. The KS test was carried out even when the omnibus was not significant. The results are shown in Table 3, which precisely reports the results of the features that were significant in at least one of the tests. 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 $\alpha = 0.05$ and an adjusted level $\alpha = 0.05/3$.

Results show that:

- 9 features are significant at the Kruskal-Wallis test, namely: the counts of hapax, nouns (NOUN), modal verbs (VM) and possessive pronouns (PP), the nouns to verb ratio, the rate of nouns (NOUNS), possessive pronouns (PP) and exclamative determiners (DE), and, above all, the te of verbs in the subjunctive mood.
- features exhibit statistically different distribution between SEVERE and MODERATE stages, at $p < 0.0.5$: noun to verb ratio, noun (NOUN) rate, modal verb count (VM), auxiliary verb count (VA), subjunctive mood rate and exclamative determiner (DE) count and rate. Noun to verb ratio, noun (NOUN) rate and the rate of verbs in subjunctive mood are significant after Bonferroni correction.
- 9 features exhibit statistically different distribution between MODERATE and MILD stages: hapax legomena count, adpositions (ADP), verbs (VERB and V POS tags), adverbs (B) and negative adverbs (BN); possessive adjective rate (AP); possessive pronoun count and rate (PP). However, after Bonferroni correction only the counts of adpositions (AP) and possessive pronouns (PP) are significant.
- as many as 22 features have statistically significant differences in their distribution between SEVERE and MILD. After the Bonferroni correction, 11 features remain significant.

Unsurprisingly, almost all the characteristics that were found to be significant (at $p < 0.05$), allow to distinguish the distributions of the SEVERE and the MILD groups, as almost all the counters in Table 3 are distributed in a significantly different way in the two groups (with slightly higher averages in the MILD than in the SEVERE). Many of these characteristics are still significantly different after Bonferroni correction ($p < 0.05/3$).

A first observation concerns a purely quantitative aspect: on average, the MILD group exhibits a greater number of words than the

Table 3
Statistically significant differences between the distribution of the linguistic features in the AIFA85 stages: SEVERE, MODERATE, MILD. The statistics values and p-values of the Kruskal-Wallis (KW) and pairwise Kolmogorov-Smirnov tests are reported for each comparison (* $p < 0.05$, ** $p < 0.01$). Bold values are significant after Bonferroni correction ($p < 0.05/3$). Tagset: the set of part-of-speech (POS) tags used to calculate the POS-based feature (UPOS/XPOS, see Tables A-2 and A-3, in the Appendix).

	Tagset	KW	Severe vs. moderate d.o.f. (60, 137)	Severe vs. mild d.o.f. (60,19)	Moderate vs. mild d.o.f. (137,19)
Words count	UPOS		0.167	0.384 *	0.257
Hapax count	UPOS	6.29 *	0.180	**0.428** **	0.334 *
Average number of words per speech turn	UPOS		0.135	0.348 *	0.230
Word-types count	UPOS		0.130	**0.395** *	0.290 .
Nouns to verbs (NOUN/VERB) ratio	UPOS	6.32 *	**0.256** **	0.298	0.273
Adjectives (ADJ) count	UPOS		0.133	0.354 *	0.277 .
Possessive adjectives (AP) rate	XPOS		0.069	0.329 .	0.323 *
Adpositions (ADP) count	UPOS		0.123	**0.395** *	**0.362** *
Adpositions (ADP) rate	UPOS		0.118	0.361 *	0.26
Coordinating conjunctions (CCONJ) count	UPOS		0.163	0.370 *	0.277
Subordinate conjunctions (SCONJ) count	UPOS		0.117	0.356 *	0.257
Nouns (NOUN) count	UPOS	6.48 *	0.194 .	0.345 *	0.188
Nouns (NOUN) rate	UPOS	8.25 *	**0.261** **	0.279	0.192
Verbs (VERB) count	UPOS		0.116	**0.423** **	0.359 *
Verbs (V) count	XPOS		0.147	**0.406** *	0.352 *
Modal verbs (VM) count	XPOS	6.18 *	0.186 *	**0.368** *	0.294 .
Auxiliary verbs (VA) count	XPOS		0.202 *	0.315 .	0.179
Rate of verbs in the subjunctive mood	UPOS	11.3 **	**0.234** *	**0.439** **	0.264
Adverbs (B) count	XPOS		0.115	**0.395** *	0.347 *
Negative adverbs (BN) count	XPOS		0.107	0.370 *	0.313 *
Clitic pronouns (PC) count	XPOS		0.132	0.351 *	0.268
Personal pronouns (PE) count	XPOS		0.115	**0.387** *	0.293 .
Possessive pronouns (PP) count	XPOS	7.26 *	0.045	**0.300** **	**0.255** *
Possessive pronouns (PP) rate	XPOS	7.26 *	0.106	**0.300** *	0.255 *
Exclamative determiner (DE) count	XPOS		0.163 *	0.167	0.088
Exclamative determiner (DE) rate	XPOS	6.76 *	0.195 *	0.333 *	0.175
Dep_dist.mean.avg	–		0.147	0.359 *	0.258

SEVERE, as demonstrated by the words_count (more specifically, a greater average_number_ of_ words_ per_ speech_turn), despite not being significantly different after Bonferroni correction.

Further features differentiating the SEVEREs from the MILDs (at $p < 0.05$) are the rate of adpositions (ADP), the rate of exclamative determiners (DE), the mean value of dependency distance, and the rate of verbs in the subjunctive mood, although only the latter is significant after the Bonferroni correction.

A subset of the counters whose distribution differs significantly (at $p < 0.05$) between the SEVERE and the MILD group can also differentiate the MODERATE from the MILD: the count of hapax, adpositions (ADP), verbs (VERBS and V), adverbs (B), negative adverbs (BN), possessive pronouns (PP). Anyway, after the Bonferroni correction ($p < 0.05/3$), only the count of adpositions (ADP) and possessive pronouns (PP) can differentiate both SEVERE from MILD and MODERATE from MILD. Also, the possessive pronouns (PP) rate turns out to play a role in the differentiating both SEVERE and MILD and MODERATE and MILD, whereas the possessive adjectives (AP) rate distinguishes only the MODERATE and MILD group (at $p < 0.05$). After the Bonferroni correction, however, PP rate and AP rate no longer distinguish the MODERATE and MILD group.

The difference between SEVERE and MODERATE (at $p < 0.05$) is made by the modal verbs (VM), auxiliary verbs (VA) and exclamative determiners (DE) counters, as well as the NOUN/VERB ratio, nouns (NOUN) rate, rate of verbs in the subjunctive mood, exclamative determiner (DE) rate. It can be seen that the rate of verbs in the subjunctive mood, the modal verbs (VM) count, and the exclamative determiner (DE) rate differ in the SEVERE group in comparison to both the MODERATE and the MILD group. After the Bonferroni correction, however, only the NOUN/VERB ratio, then nouns (NOUN) rate and the rate of verbs in the subjunctive mood can distinguish the SEVERE and the MODERATE group. The latter feature can also distinguish the MODERATE and the MILD group.

As for the classification task, all the available features were calculated, regardless of which features were found to be significant. This ensured that any favorable interaction between significant and non-significant features would also be taken into account. The most performing variant of the classifiers illustrated in Section 4.2.2.1 adopted multivariate normal (MVN) modeling of the features distribution, and Forward Backward selection (FBsel) as feature selection strategy. The resulting evaluation metrics are shown in Table 4. It reports macro F1, precision and recall scores both for 3-class and for each single 2-class classification task.

Most of the literature considers only a 2-class classification scenario (e.g., AD vs. healthy, or AD vs. MCI), so any comparison can only be made mainly on this basis. In this sense, according to our results, the highest accuracy is obtained in classifying SEVERE vs. MILD. The F1 score is not far from Calzà et al.'s (2021), which also includes other results from the literature. However, we should note that even a direct match in this case would not be particularly valid, since the researchers focused on the classification of MCI and healthy controls (F1 = 0.7045) and their set was larger, including, for example, audio features.

5.2.2. *Classification via language models and perplexity*

This experiment was designed to test the discriminative power of both language models and perplexity scores. We assessed two simple categorization approaches to discriminate between patients grouped by the AIFA85 classification according to their MMSE score.

Firstly, we tried to categorize the transcripts into the three main categories (mild, moderate, and severe) by relying on the conversation transcripts. The results were then compared against the classes based on the MMSE score: SEVERE [0–9], MODERATE [10–20] and MILD [21–26]. In this setting the categorization was performed by BERT and was mapped onto the three labels through a classifier using the Sparse Categorical Cross-Entropy as the loss function. The BERT multilingual uncased base model was augmented by including the ParlaTO corpus conversations of those aged over 60. The experiments were performed in 10-fold cross-validation.

The results are illustrated in Table 5: the F1 scores across the three classes are 0.225 for SEVERE, 0.717 for MODERATE and 0.053 for the MILD. The final macro-average F1 was 0.341. Such very different F1 scores, leading to a quite low macro value, may be partly explained by the reduced size and imbalance of the experimental data. If we consider the rightmost column of the table ('support'), we observe that the MODERATE class has more than double the elements of the SEVERE class, which in turn has three times as many elements as the MILD class. In this setting, it is not surprising that our classifier obtains its best results from the MODERATE class.

To explore the effect of the imbalance of the classes, we re-ran this experiment by employing balanced classes (19 support items per class), sized on the basis of the MILD class, the smallest class in the subset of the Anchise Corpus, resulting in a 0.40 macro-averaged F1. This significant improvement strongly suggests that the imbalance skewed the original results, but still does not allow us to be conclusive about the models' ability to discriminate.

To further explore the data at hand we devised a new experimental setting, where we categorized classes into pairs: SEVERE-MILD, SEVERE-MODERATE, MODERATE-MILD; and then compared the discriminative features of a perplexity-based classifier with that of a classifier relying on BERT. The calculation of perplexity scores was based on a usage of perplexity similar to previous work in the literature (Fritsch et al., 2019; Cohen and Pakhomov, 2020; Colla et al. 2022). Three different language models were employed (one for each of the classes compared: LM_{Sev}, LM_{Mod}, LM_{Mil}); for each pair of classes c_1 and c_2 (classes are sorted in such a way that c_1 is assumed to be more severe than c_2), the PPL^p score for a given patient p was computed as

$$PPL^p = PPL(LM_{c_1}, t^p) - PPL(LM_{c_2}, t^p).$$

For each class c, $PPL_{c \mid c \in \{c_1,c_2\}}$ is averaged over PPL^p scores, and the class for patient $p \in c$ is predicted as

$$class(p) = argmin_{x \in \{c_1,\ c_2\}} \left| PPL^p - PPL_{x \mid x \in \{c_1,c_2\}} \right|.$$

Each subject transcript $t^{p \mid p \in c}$ is dropped both while training LM_c, and when computing PPL_c.

The results are presented in Table 6: the macro-average of the F1 figures rises to 0.6065. Compared to the results obtained by

Table 4
Results of the classification by means of digital linguistic biomarkers. Macro F1, precision and recall.

	F1	Precision	Recall
Severe vs. moderate vs. mild	0.608	0.590	0.661
Severe vs. moderate	0.724	0.719	0.738
Severe vs. mild	0.789	0.769	0.824
Moderate vs. mild	0.698	0.669	0.789

Table 5
Results for the classification task. We reported precision, recall, F1-measure accuracy and support for each class.

Class	Precision	Recall	F1	Accuracy	Support
Severe	0.345	0.167	0.225	0.681	60
Moderate	0.630	0.832	0.717	0.583	137
Mild	0.167	0.053	0.080	0.894	19

applying the BERT-based categorization (Fig. 4), we observe that the scores calculated with the perplexity mostly outdo those calculated by BERT, which only reach a 0.483 macro-averaged F1 (Table A-6). We should also note that the BERT-based categorization had been improved with the addition of the over-60 s age group from the ParlaTO corpus.

We argue that the advantage of the perplexity-based approach is largely due to the fact that the BERT-based categorization was critically affected by the size of the support sets: we had 60 transcripts for SEVERE, 137 for MODERATE and only 19 for MILD. We can also conclude that the classifier employing the perplexity scores mostly succeeded in building a more informed representation on the language.

5.2.3. Classification within sentiment and emotion analysis

Table 7 illustrates the distribution of the positive, negative, and neutral conversations, within each AIFA85 stage. The same tendency found with the correlation analysis is also present between the SEVERE to MODERATE to MILD stages. In all cases, as the MMSE increases, there is an increase of negative sentiment, while the positive sentiment decreases. However, it should be noted that Fisher's Exact Test for Count Data reveals a statistically significant difference only for the positive sentiment and only between SEVERE and MILD (p-value = 0.0328), that, however, disappears after the Bonferroni correction ($p < 0.05/3$). That said, in both stages there is always a tendency for negative sentiment to prevail over others.

As for emotion analysis, Table 8 shows the emotion frequencies per 100 speech turns within each AIFA85 stage. According to the Fisher's Exact Test for Count Data:

- the proportion of joy is statistically significant in the comparison of SEVERE vs. MODERATE ($p < 0.001$), even after Bonferroni correction ($p < 0.05/3$); significant at $p < 0.05$ in comparison of SEVERE vs. MILD ($p = 0.0194$), while no statistically significant difference was found between MODERATE and MILD classes.
- the proportion of sadness is statistically significant in the comparison of SEVERE vs. MODERATE ($p < 0.001$) and SEVERE vs. MILD ($p = 0.00377$), even after Bonferroni correction, while no statistically significant difference was found between MODERATE and MILD classes.
- both for ANGER and FEAR, no statistically significant difference was found at all.

Thus, there are proportionally more joyful (and fewer "sad") speech turns in the SEVERE class than in the MODERATE, and in the SEVERE than in the MILD class.

As for the emotion rates, the results with the Kolmogorov-Smirnov test are shown in Table 9. Among the negative emotions, the anger rate tends to be statistically significant, at $p < 0.05$, in the comparison between SEVERE and MILD, and between MODERATE and MILD, but no longer after the Bonferroni correction ($p < 0.05/3$); the sadness rate is significantly different only in the SEVERE vs. MODERATE comparison, even after the Bonferroni correction, while the fear rate difference is insignificant. The joy rate significantly changes only in the comparison between SEVERE and MILD, but only at $p < 0.05$.

The AIFA85 categorization allows us to refine the trend already found in the correlation analysis. In particular, shifting from

Table 6
Results for the categorization experiment, using the PPL scores: precision, recall, F1-measure accuracy and support for each binary classifier tested.

Pair	Class	Precision	Recall	F1	Accuracy	Support
Severe vs Mild	Severe	0.792	0.7	0.743	0.633	60
	Mild	0.308	0.421	0.356		19
Severe vs Moderate	Severe	0.59	0.6	0.595	0.751	60
	Moderate	0.824	0.818	0.821		137
Moderate vs Mild	Moderate	0.931	0.693	0.795	0.686	137
	Mild	0.222	0.632	0.329		19

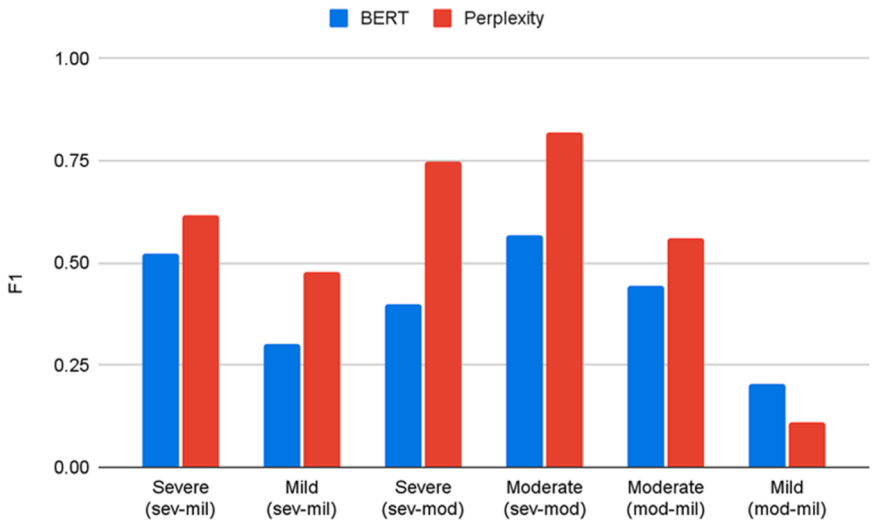


Fig. 4. Comparison between the F1 scores obtained by employing the BERT-based and perplexity-based categorization. Each class appears twice, as it is involved in two binary comparisons.

Table 7
Distribution of positive, negative, and neutral conversations across the AIFA85 stages.

	Mild	Moderate	Severe
Negative	68.4 %	59.9 %	53.3 %
Positive	15.8 %	29.9 %	43.3 %
Neutral	15.8 %	10.2 %	3.3 %

Table 8
Emotions/turn of speech across the AIFA85 stages.

	Mild	Moderate	Severe
Anger	14.0 %	12.3 %	12.6 %
Fear	2.8 %	2.9 %	2.9 %
Joy	40.9 %	41.4 %	47.3 %
Sadness	42.3 %	43.4 %	37.3 %

Table 9
Statistical comparison with the Kolmogorov-Smirnov test of emotions. p-value are reported (* $p < 0.05$, ** $p < 0.01$). Bold values are significant after Bonferroni correction ($p < 0.05/3$). D.o.f.: degrees of freedom.

	Severe vs. moderate d.o.f. (60, 137)	Severe vs. mild d.o.f. (60, 19)	Moderate vs. mild d.o.f. (137, 19)
feelit_anger_rate	0.111	0.381 *	0.327 *
feelit_fear_rate	0.061	0.192	0.168
feelit_joy_rate	0.195 .	0.381 *	0.188
feelit_sadness_rate	**0.262** **	0.256	0.211

SEVERE to MODERATE, the increase of negative sentiments is due mainly to sadness, while from MODERATE to MILD there is only a weak significant increase of anger. The decrease in the joyful expressions, at session level, is most evident between the extremes, SEVERE and MILD, although not after the Bonferroni correction.

These findings may reflect the patient's emotional reaction to awareness of cognitive impairment. In the initial phase the patient reacts with anger and sadness; in the advanced stage, when the subject has lost reference to the current world and lives in her/his own possible world without realizing it, s/he rediscovers a sort of serenity based precisely on the loss of awareness of her/his own neurocognitive deficits.

5.3. Discussion

The in-depth analysis of the transcribed conversations was challenging in many aspects. The Anchise Corpus was not created with any research criteria in mind, and the conversations themselves were freely elicited in ecologically valid conditions. The analysis was further hampered given the sporadic nature of further socio-linguistic and specific diagnostic information. However, despite these extremely unusual premises, the automatic analysis of the Anchise Corpus has offered precious qualitative and quantitative data on language production in dementia.

The adoption of automatic tools and the focus on different types of information (in line with our first sub-objective, see Section 3), have allowed different perspectives of corpus analysis. No tool or type of information under examination proved to work clearly better than others. Rather, all tools turned out to produce better results in classification tasks that considered AIFA85 severity levels rather than in correlating transcribed speech characteristics with individual MMSE scores. This type of correlation may be more reasonable to look for, given the intersubjective variability and the fact that the MMSE test is not specific for language or to test the subject's emotional state and sentiment.

That said, a correlation between DLBs of oral production and MMSE scores was found (in line with our subgoal 2a), albeit weak. This could also be due to the task and related speech style we have investigated (in line with results in Beltrami et al. 2016). Nevertheless, the significant correlations that we found are consistent with those in the literature, such as with Hernández-Domínguez et al. (2018), who also found the MMSE to be positively correlated with the hapax legomena count.

Overall, the highest accuracy is obtained in classifying the SEVERE vs. the MILD group. However, results suggest that different sets of features tend to discriminate between the MILD group from the others, and, on the other hand, the SEVERE group from the others. As for the former case, the count of hapax, possessive pronouns, verbs (VERB and V), and, to a less significant extent, adpositions, and adverbs, as well as the possessive pronouns rate, tend to discriminate MILD from both SEVERE and MODERATE, being slightly higher in the former. On the other hand, the rate of verbs in the subjective mood and, to a less significant extent, the modal verbs count and the exclamative determiner rate tend to discriminate the SEVERE group from both the MILD and the MODERATE, being slightly lower in the former.

In light of the findings related to the linguistic features, it seems reasonable to infer the following.

1) Vocabulary impoverishment. The correlation analysis results reported in Table 2 show that as the MMSE score decreases, the number of hapax legomena also decreases. This trend is not necessarily due to a decrease in conversation duration (word count), because no significant trend of word count was detected in the correlation analysis. In particular, Table 3 shows that the number of words is distributed differently only between the SEVERE and MILD groups (with higher values for MILD), while the number of hapax legomena is distributed differently not only between SEVERE and MILD but also between MODERATE and MILD (with higher values for MILD). This can be interpreted as the sign of a progressive impoverishment of the lexicon as cognitive impairment advances.
2) Simplified use of the verbs. Overall, it seems that the use of verbs becomes more simplified as the MMSE decreases. As a first indicator there is a general tendency, as the MMSE score decreases, to prefer the present tense (*rate of verbs in the present tense*) to the detriment of the past tense (*rate of verbs in the past tense*). Likewise, a general tendency is found, as the MMSE score decreases, to prefer the finite verbs (*finite verb rate*) and less subjunctive verbs, modal verbs and participle and gerund verb forms (see Table 2). Furthermore, a greater use of auxiliary verbs emerges from the SEVERE to the MODERATE group (with slightly higher averages in MODERATE). Consistently, results concerning verbs reported in Table 3 seem to significantly differentiate the MILD group from the others (see the slightly higher averages of VERB and V in MILDs in behind the differences in Table 3, and discussion above); further, results concerning a more limited use of subjunctive verbs in the SEVERE group in comparison to the others point in the same direction.
3) Informality. We find a negative correlation of interjection rate (Table 2), that could suggest an increase in the adoption of a more informal register (Mereu and Vietti, 2021). Further, it appears that the exclamatory use of exclamative determiners (DE) increases among people with progressively lower MMSE. We might hypothesize here a tendency for subjects with lower MMSE scores to adopt a more informal register (Tables 2, 3).
4) Anomia. We also found that the rate of nouns decreases as the MMSE score decreases, which is consistent with anomia, the first language disorder to appear in the ILDW, especially with AD. The ratio between nouns and verbs is also in line with this result, although it can only distinguish between SEVERE and MODERATE (with lower values in the SEVERE group).

Correlations between MMSE and perplexity scores (see again our subgoal 2a, see Section 3) also turned out to be weak, despite promising results obtained in previous work on picture descriptions in AD (Colla et al., 2022). However, besides differences in the number of transcripts, the corpus was particularly raw and uncurated, as mentioned above. Even though the results did not present high correlations with the MMSE scores, perplexity proved to be a valid approach, mostly outperforming the approaches based on BERT. The perplexity score is essentially an information-theoretic measure designed to evaluate the coherence of a given text sequence based on a given LM. It thus requires further refinement to gather fine-tuned data coming from naturalistic settings such as that of the Anchise Corpus. Additionally, comparisons should be made to explore potential connections with standard linguistic features that, to our knowledge, have never been explored in the literature.

Above all the lack of any strong correlation is in line with the intersubjective variability observed in the clinical practice, as well as with the intrinsic non-linguistic nature of the MMSE assessment. This fact entails that even though more severe impairment should in principle be associated with lower MMSE scores, nonetheless consistent differences may be observed. For example, a patient suffering from Alzheimer's may even be assigned a full MMSE score.[6] These findings take on more validity when we consider the fact that language impairments in people with dementia affect both cohesion and coherence. Previous studies have shown that coherence declines before cohesion, that is, the semantic connections between words deteriorate first, and only later on, are morphosyntactic connections affected (Boschi et al., 2017).

As for sentiment and emotions, the results show that the rate of negative emotions decreases as the MMSE decreases. Considering once again the naturalistic condition in which patients are free to talk about themselves in a very spontaneous manner, this 'joyful' result is consistent with a gradual reduction in awareness of reality. Indeed, ILWD seem to recall positive emotions more easily with progression of the disease. Further, we hypothesize that this effect is in line with dementia patients who are often also affected by depression, with depression possibly masking the diagnosis of dementia itself. So, as the disease progresses, patients become less connected to reality, which also includes losing their depression. This conclusion suggests that sentiment and emotion analysis could help in improving the classification by tools relying on linguistic information only, as well as being something to consider when investigating depression itself. In this sense, having obtained statistically significant, albeit not strong, correlation values using pre-trained computational models does suggest the validity of further analysis of the emotional sphere using ecologically valid conversations with other computational models, or taken from other naturalistic settings.

As for distinguishing different stages of impairment by reference to the MMSE (subgoal 2b, Section 3), the tools developed to investigate lexical and morphosyntactic features did reveal interesting results. For example, for the binary categorization of SEVERE vs. MILD class, a F1 score close to 0.79 was achieved, complemented by a macro-averaged F1 = 0.61 in the 3-class categorization.

On the other hand, state-of-the-art approaches based on language models, either based directly on BERT models or on perplexity, calculated by models based on GPT-2 only attain limited correlation with MMSE scores. The results were in the order of a 0.26 inverse correlation using Pearson *r*, and only a limited categorization accuracy of 0.40 macro-F1 for the multi-label categorization using BERT-based classifiers, and 0.60 macro-F1 for the binary categorization. Evidence was found that in this setting the explicit knowledge of human experts ensures higher accuracy with respect to neural, feature-agnostic approaches.

A number of factors seem to have undermined the accuracy of classifiers employing LMs. First there was data imbalance. Though the Anchise Corpus is larger than other corpora, it is still below what is necessary to fully inform language models on specific application domains. Furthermore, the naturalistic and free nature of the conversations created issues as well as the weak connection between linguistic production and MMSE score. Also, there was no control group. Nevertheless, the perplexity metrics provided improved results with respect to the BERT-based classifiers (0.60 vs. 0.49 macro-F1 score). This seems to suggest that, for limited and imbalanced data samples, perplexity is preferable to more standard CLS-based categorization approaches. To compare our system against the literature, we calculated the correlation between perplexity scores and MMSE scores in the Pitt Corpus (focusing on the description of the cookie-theft picture), obtaining a correlation of 0.73. This result is in line with Fritsch et al. (2019), who obtained a correlation of 0.66 (using Pearson coefficients). Interestingly enough, Fritsch and colleagues (2019) report a 0.66 correlation for the whole data, but poorer figures for the two classes taken in isolation: 0.43 for AD and 0.11 for Healthy Controls. Such lower correlations are perhaps best understood if one considers the control class, where many perplexity scores need to be mapped onto a few MMSE scores (just those ranging from 24 to 30), resulting in a low correlation. This result strengthens the argument that our own results on the Anchise Corpus can be improved by adding a control class.

Finally, in line with the third subgoal stated in Section 3, the Anchise 2022 Corpus was analyzed per se, as a precious source of information regarding naturalistic dialogical speech in people afflicted with dementia. The analysis shows that also in such a kind of dialogues the disease progression corresponds to the impoverishment of the vocabulary, anomia, and a simplified use of the verbs; further, a tendency to the use of a more informal register is observed. Moreover, the coherence tends to diminish also in dialogical speech (as shown by the analysis of perplexity), while the general sentiment and emotions tend to improve in the direction of a "joyful" or less sad state. This is consistent with a tendency to gradually focus on a present (tens of verbs) characterized by a reduced awareness of reality. These features may be more clearly observed in dialogues collected within a naturalistic setting in which subjects are free to express themselves, as in the Anchise 2022 corpus, rather than in more traditional experimental setups.

6. Conclusions and future work

Our study provided insights into the language production of ILWD involved in conversation in naturalistic conditions. It investigated the ability of text-based sources of information and of NLP tools to correlate these characteristics with the degree of severity of the cognitive impairment, as indicated by their MMSE score. More than 200 transcribed conversations from the Anchise 2022 Corpus were submitted to a wide range of automatic tools, from more traditional morpho-syntactic features extraction and categorization to the more recent BERT-based classification and perplexity measurement, to text-based sentiment and emotion analysis. This is the first experiment on a large corpus of spontaneous dialogues in Italian.

We emphasize the ecological validity of the analysis of speech performed in the present study, which differs from most existing literature, which is focused on controlled speech experiments. Additionally, our experiments targeted a dataset composed of over 200

[6] This fact can be easily verified by extracting the MMSE scores associated with the subjects in the AD class from the Pitt Corpus (Becker et al., 1994).

conversations, that is a noteworthy dataset, especially because it is in Italian, which is often underrepresented. Furthermore, this research has introduced advanced linguistic models for analyzing perplexity, sentiment (polarity), and emotion—a departure from traditional approaches in this field.

Integrating various profiles of analysis, such as DLBs perplexity, emotion, and sentiment analysis, has been proved to be effective in offering a wider picture of linguistic and communication deficits, as well as offering more precise data on the progression of dementia.

Regarding future research, the results suggest the need to shift the focus from the search for general classifiers to the identification of language impairments at each stage of the disease. Like other fields of medical sciences, it would be beneficial to study approaches specifically designed to deal with a specific subject and a specific stage of impairment. The identification of specific issues in language production can help guide personalized supportive rehabilitative treatment, particularly in speech therapy. It can also provide valuable for professional caregivers and family members and foster a more positive approach to ILWD. Further, detailed subject-tailored analyzes should also be a boost for diagnosis, making it more specific, and help in diagnosing dementia earlier. We can reasonably hope for the future that more nuanced analysis techniques will further improve accuracy in the description of the characteristics of naturalistic language production in relation to the MMSE score.

CRediT authorship contribution statement

Francesco Sigona: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Daniele P. Radicioni:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization. **Barbara Gili Fivela:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Davide Colla:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Matteo Delsanto:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Enrico Mensa:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Andrea Bolioli:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Pietro Vigorelli:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Francesco Sigona reports financial support was provided by Italian Ministry of University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

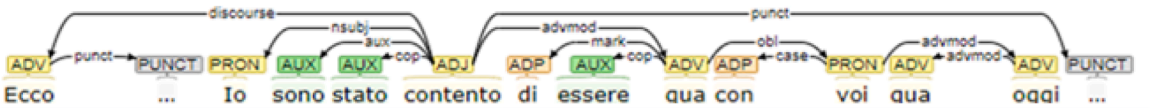
Data availability

Data will be made available on request.

Funding

This work was partially funded by the Italian Ministry of University and Research, D.M. 1062/2021 (within the innovation topics - Actions IV.4 - scientific sector L-LIN/01).

Appendix A



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	

Fig. A-1. 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 (Qi et al., 2020) for the same sentence, in the CONLL-U format (Nivre et al., 2016).

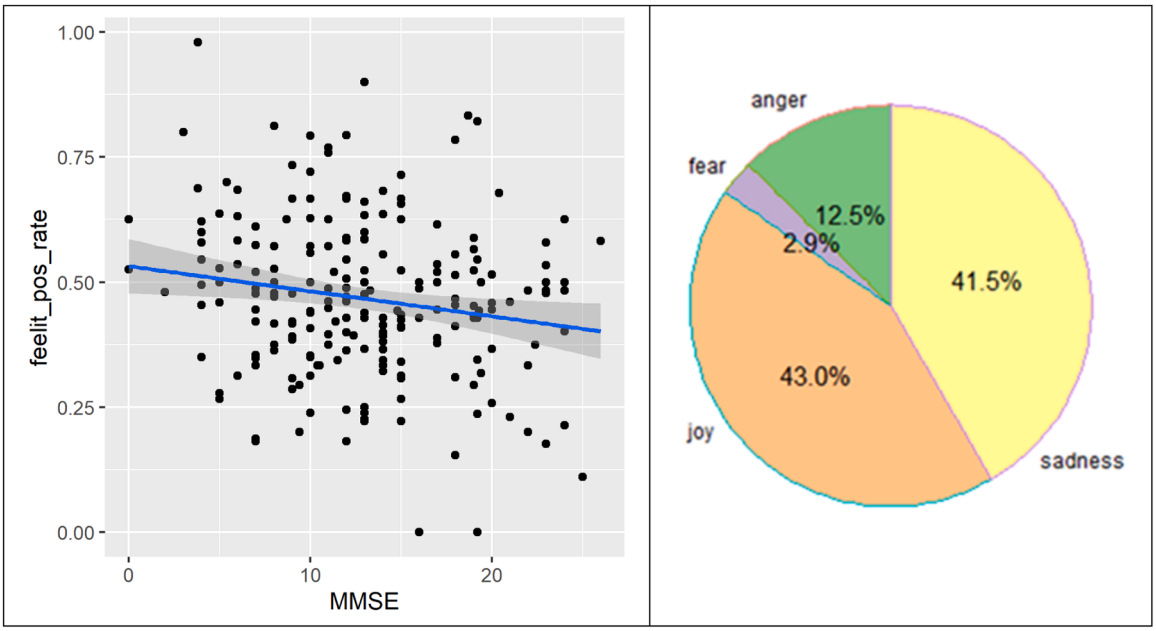


Fig. A-2. On the left: positive sentiment rate (*pos_rate*) vs. mini-mental state examination score (MMSE): scatter plot and trend line. The linear fit explains only a low portion of the whole variance ($R^2 = 0.03$). However, the trend ($\beta = -0.005$) is statistically significant ($p = 0.012$ *) and shows that the positive sentiment slightly decreases as the MMSE increases. On the right: proportions of emotions (turns of speech) over the Corpus.

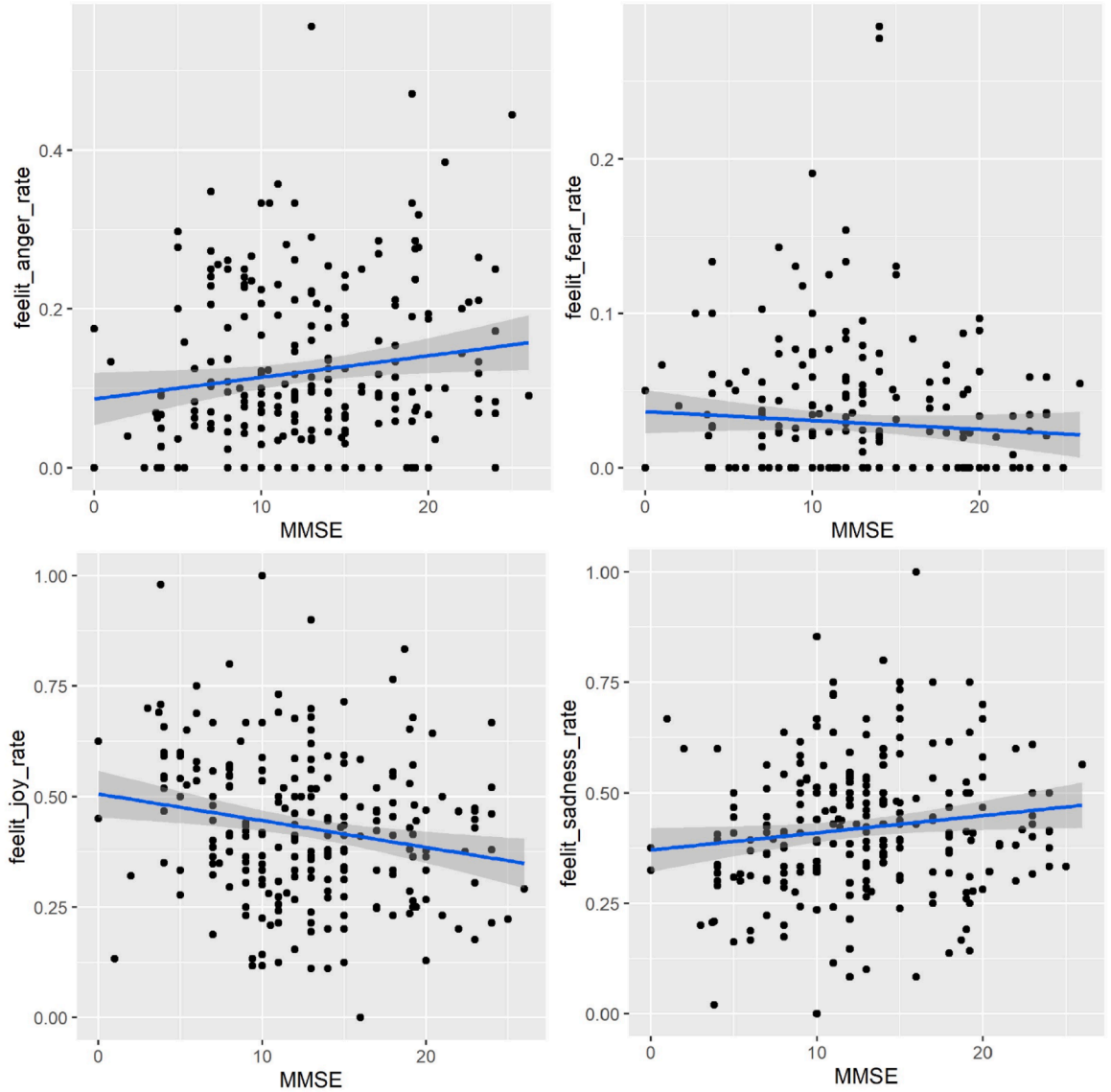


Fig. A-3. Scatterplot and linear trends of anger, fear, joy and sadness rates at conversation level, against mini-mental state examination score (MMSE). The emotion of fear is scarcely present, and has no correlation with MMSE, while the other emotions show significant correlation both for Pearson and Spearman coefficients. As MMSE increases, joy slightly decreases (β = -7.4, Pearson c.c. = -0.22, $p < 0.001$***), while anger (β = 8.4, Pearson c.c. = 0.15, $p = 0.0176$ *) and sadness (β = 5.7, Pearson c.c. = 0.16, $p = 0.0161$ *) slightly decreases.

Table A-1
Some of the attributes calculated by the Stanza tool (Qi et al., 2020) for each detected token in a sentence. In bold font, the data used in this study.

ID	A word index. Usually an integer number starting at 1 for each new sentence.
FORM	**Word form or punctuation symbol**
LEMMA	Lemma or stem of word form
UPOS	**Universal part-of-speech tag**
XPOS	**Language-specific part-of-speech tag; underscore if not available.**
FEATS	**List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available**
HEAD	**Head of the current word, which is either a value of ID or zero (0)**
DEPREL	Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.

Table A-2
Universal POS tags.

UPOS tag	Description	UPOS tag	Description
ADJ	Adjective	NUM	Numeral
ADP	Adposition	PART	Particle
ADV	Adverb	PRON	Pronoun
AUX	Auxiliary	PROPN	Proper noun
CCONJ	Coordinating conjunction	PUNCT	Punctuation
DET	Determiner	SCONJ	Subordinating conjunction
INTJ	Interjection	SYM	Symbol
NOUN	Noun	VERB	Verb
		X	Other

Table A-3
XPOS tags for Italian.

XPOS tag	Description	XPOS tag	Description
A	Adjective	N	Cardinal number
AP	Possessive adjective	NO	Ordinal number
B	Adverb	PC	Clitic pronoun
BN	Negative adverb	PD	Demonstrative pronoun
CC	Coordinate conjunction	PE	Personal pronoun
CS	Subordinate conjunction	PI	Indefinite pronoun
DD	Demonstrative determiner	PP	Possessive pronoun
DE	Exclamative determiner	PQ	Interrogative pronoun
DI	Indefinite determiner	PR	Relative pronoun
DR	Relative determiner	RD	Determinative article
DQ	Interrogative determiner	RI	Indeterminative article
E	Preposition	S	Common noun
EA	Articulated preposition	SP	Proper noun
FB	Balanced punctuation	SW	Foreign noun
FC	Clause boundary punctuation	T	Predeterminer
FF	Comma	V	Verb
FS	Sentence boundary punctuation	VA	Auxiliary verb
I	Interjection	VM	Modal verb
X	Residual class: it includes formulae, unclassified words, alphabetic symbols and the like		

Table A-4
Digital linguistic biomarkers extracted session-by-session after Natural Language Processing. In addition to the UPOS tags, also the XPOS tag rate (number of occurrences per 100 words) are used in the present work. Notes: (*) open-class words have UPOS tags: "ADJ", "ADV", "INTJ", "NOUN", "PROPN", "VERB"; (**) closed-class words have UPOS tags: "ADP", "AUX", "CCONJ", "DET", "NUM", "PART", "PRON", "SCONJ".

Group	Feature name	Description
Lexical	content_density	The ratio of open-class words (*) to open + closed class words (**) (Roark et al., 2011)
	open_to_close_class	The ratio of open-class words to closed-class words (**)
	tokens_to_turns	The ratio between the number of tokens and the number of 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 letters of the words
Semantic & Lexical richness	num_hapax_legomena	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, that has been used in stylometric analyzes of text and is 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 {0.172, 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 v_1 is the number of hapax, and V is the number of types
Lexical / syntactic	adj_perc	Adjectives (ADJ) rate per 100 words
	adp_perc	Adpositions (ADP) rate per 100 words
	adv_perc	Adverbs (ADV) rate per 100 words
	aux_perc	Auxiliary (AUX) rate per 100 words
	cconj_perc	Coordinate conjunctions (CCONJ) rate per 100 words
	sconj_perc	Subordinate conjunctions (SCONJ) rate per 100 words

(*continued on next page*)

Table A-4 (*continued*)

Group	Feature name	Description
	sconj_to_cconj	the ratio between the number of subordinate and the number of coordinate conjunctions (sentences)
	noun_perc	Nouns (NOUN) rate per 100 words
	nouns_to_verbs	The ratio between the number of nouns and the number of verbs. This index is also referred to as "reference rate to reality" (Vigorelli, 2004)
	pron_perc	Pronouns (PRON) rate per 100 words
	pronouns_to_nouns	The ratio between the number of pronouns and the number of nouns
	intj_perc	Interjections (INTJ) rate per 100 words
	propn_perc	Proper nouns (PROPN) rate per 100 words
	verb_perc	Verbs (VERB) rate per 100 words
	x_perc	The rate of words that could not be assigned to any other category (X) per 100 words.
Syntactic	dependency distance	Dependency distance (Roark et al., 2007,2011). Mean, standard deviation and maximum values have been considered across each sentence.
	max_depth	Maximum structure depth (mean, standard deviation, maximum across each sentence)
	szmrecsanyi	A syntactic complexity measure by Szmrecśanyi, 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	Number of participle and gerund verbal forms
	reduced_sentences_to_verbs	Number of participle and gerund verbal forms divided by the number of 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

Table A-5
Linear fit and correlation coefficients of anger, fear, joy and sadness rates against mini-mental state examination score (MMSE). Statistically significant features ($\alpha = 0.05$) are highlighted in bold font ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Abbreviations: beta, the coefficient of the linear fit; t, the value of the statistic, with 236 degrees of freedom; CI, confidence interval of the estimation of beta; R2, R-squared, the fraction of the explained variance; r-Pearson and ρ-Spearman, correlation coefficients).

Feature	Beta	t (236)	CI	R2	r-Pearson	ρ-Spearman
Anger	**8.4346**	**2.3911**	**[1.4851, 15.384]**	**0.0237**	**0.1538** *	**0.1297***
Fear	-8.1833	0.9731	[-24.7513, 8.3848]	0.0040	-0.0632	-0.0684
Joy	**-7.4205**	**−3.4814**	**[-11.6196, -3.2213]**	**0.0488**	**-0.2210** ***	**-0.2223*****
Sadness	**5.6987**	**2.4252**	**[1.0694, 10.328]**	**0.0243**	**0.1559***	**0.1697** **

Table A-6
Results for experiment 2 where categorization was computed by employing the BERT-based classifier. We reported Precision, Recall, F1-measure Accuracy and Support for each class pairs combination.

Pair	Class	Precision	Recall	F1	Accuracy	Support
Severe vs mild	Severe	0.757	0.933	0.836	0.722	60
	Mild	0.2	0.053	0.083	0.722	19
Severe vs moderate	Severe	0.366	0.250	0.297	0.640	60
	Moderate	0.712	0.810	0.758	0.640	137
Moderate vs mild	Moderate	0.876	0.978	0.924	0.859	137
	Mild	0.000	0.000	0.000	0.859	19

References

Altun, H., Polat, G., 2009. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst. Appl. 36 (4), 8197–8203. https://doi.org/10.1016/J.ESWA.2008.10.005.

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). Cochrane Database Syst. Rev. 2021 (7), CD010783 https://doi.org/10.1002/14651858.CD010783.pub3.

Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J., 2020. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. In: Proceedings of Interspeech 2020 (Shanghai), pp. 2167–2171. https://doi.org/10.21437/Interspeech.2020-2557.

Banovic, S., Zunic, L.J., Sinanovic, O., 2018. Communication Difficulties as a Result of Dementia. Mater. Sociomed. 30, 221–224. https://doi.org/10.5455/msm.2018.30.221-224, 2018.

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. Arch. Neurol. 51 (6), 585–594, 1994.

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). Portorož, Slovenia. European Language Resources Association (ELRA), pp. 2086–2093 pages.

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? Front. Aging Neurosci. 10, 369. https://doi.org/10.3389/fnagi.2018.00369, 2018.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. Noise Reduction in Speech Processing. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5. Springer Topics in Signal Processing, vol 2.

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. (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020: Bologna, Italy, March 1-3, 2021, Torino. Accademia University Press, pp. 51–57. https://doi.org/10.4000/books.aaccademia.8260.

Bernard, B., Goldman, J.G., 2010. MMSE - mini-mental state examination. In: Kompoliti, K., Verhagen Metman, L. (Eds.), Encyclopedia of Movement Disorders. Academic Press, pp. 187–189. https://doi.org/10.1016/B978-0-12-374105-9.00186-6, 2010.

Bianchi, F., Nozza, D., Hovy, D., 2021. FEEL-IT: emotion and sentiment classification for the Italian language. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 76–83 pagesOnline. Association for Computational Linguistics.

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. December 2-4, 2020. ISBN 978-88-3379-286-6.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., Cappa, S.F., 2017. Connected speech in neurodegenerative language disorders: a review. Front. Psychol. 8, 269. https://doi.org/10.3389/fpsyg.2017.00269, 2017.

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. Behav. Sci. 12, 339. https://doi.org/10.3390/bs12090339, 2022.

Calzà, L., Gagliardi, G., Rossini Favretti, R., Tamburini, F., 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput. Speech Lang. 65, 101–113. https://doi.org/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***.

Cohen, T., Pakhomov, S., 2020. A tale of two perplexities: sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1946–1957. https://doi.org/10.18653/v1/2020.acl-main.176, 2020.

Colla, D., Mensa, E., Radicioni, D.P., 2020. LESSLEX: linking multilingual embeddings to sense representations of lexical items. Comput. Linguist. 46 (2), 289–333. https://doi.org/10.1162/coli_a_00375 pages.

Colla, D., Delsanto, M., Agosto, M., Vitiello, B., Radicioni, D.P., 2022. Semantic coherence markers: the contribution of perplexity metrics. Artif. Intell. Med. 134, 102393 https://doi.org/10.1016/j.artmed.2022.102393.

Colla, D., Delsanto, M., Radicioni, D.P., 2023. Semantic coherence dataset: speech transcripts. Data Brief 46, 108799. https://doi.org/10.1016/j.dib.2022.108799.

Creavin, S.T., Wisniewski, S., Noel-Storr, A.H., Trevelyan, C.M., Hampton, T., Rayment, D., et al., 2016. Mini-mental state examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. Cochrane Database Syst. Rev. 2016 (1), CD011145 https://doi.org/10.1002/14651858.CD011145.pub2.

Cummings, J.L., Mega, M.S., Gray, K., Rosemberg-Thompson, S., Gornbein, T., 1994. The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. Neurology. 44, 2308–2314. https://doi.org/10.1212/wnl.44.12.2308, 1994.

DeJong, R., Osterlund, O.W., Roy, G.W., 1989. Measurement of quality-of-life changes in patients with Alzheimer's disease. Clin. Ther. 11 (4), 545–554, 1989 Jul-AugPMID: 2776169.

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. J. Alzheimer's Dis. 78 (4), 1547–1574. https://doi.org/10.3233/JAD-200888.

De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., Guerini, M. (2020). Geppetto carves Italian into a language model. arXiv preprint arXiv:2004.14253.

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. Alz. Res. Therapy 11, 21. https://doi.org/10.1186/s13195-019-0474-3 (2019).

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. Dement. Neurocogn. Disord. 20 (4), 89–98. https://doi.org/10.12779/dnd.2021.20.4.89, 2021 Oct.

Devlin, J., Chang M.W., Lee K., Toutanova K. (2018) Bert: pre-training of deep bidirectional transformers for language understanding. 2018, *arXiv preprint* arXiv: 1810.04805.

Online 2021 Dovetto, F.M., Guida, A., Pagliaro, A.C., Guardasci, R., Raggio, L., Sorrentino, A., Trillocco, S., 2022. Corpora di Italiano parlato patologico dell'età adulta e senile (a cura di). In: Cresti, E., Moneglia, M. (Eds.), *Corpora e Studi Linguistici*. Atti del LIV Congresso Della Società di Linguistica Italiana. Officinaventuno, Milano, pp. 165–177.

Ekman, P., 1992. An argument for basic emotions. Cogn. Emot. 6 (3–4), 169–200. https://doi.org/10.1080/02699939208411068.

Espinoza-Cuadros, F., Garcia-Zamora, M.A., Torres-Boza, D., Ferrer-Riesgo, C.A., Montero-Benavides, A., Gonzalez Moreira, E., Hernández-Gomez, L.A., 2014. A spoken language database for research on moderate cognitive impairment: design and preliminary analysis. Advances in Speech and Language Technologies For Iberian Languages. Springer, Cham. https://doi.org/10.1007/978-3-319-13623-3_23. Lecture Notes in Computer Science(2014), vol 8854.

Ferris, S.H., Farlow, M., 2013. Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. Clin. Interv. Aging 8, 1007–1014. https://doi.org/10.2147/CIA.S39959, 2013.

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. Aphasiology. 34 (6), 723–755. https://doi.org/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. J. Psychiatric Res. 12, 189–198. https://doi.org/10.1016/0022-3956(75)90026-6, 1975.

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, pp. 5841–5845. https://doi.org/10.1109/ICASSP.2019.8682690, 2019.

Gagliardi, G., 2023. Natural language processing techniques for studying language in pathological ageing: a scoping review. Int. J. Lang. Commun. Disord. https://doi.org/10.1111/1460-6984.12870.

Gagliardi, G., Tamburini, F., 2021. Linguistic biomarkers for the detection of mild cognitive impairment. 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 Thirteenth Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association, pp. 5234–5242 pages.

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. Cortex 55, 122–129.

Gélinas, I., Gauthier, L., McIntyre, M., Gauthier, S., 1999. Development of a functional measure for persons with Alzheimer's disease: the disability assessment for dementia. Am. J. Occup. Ther. 53, 471–481. https://doi.org/10.5014/ajot.53.5.471, 1999.

Goldberg, Y., 2017. Neural network methods for natural language processing. Synth. Lect. Hum. Lang. Technol. 10 (1), 1–309. Springer.

Goodglass, H., Kaplan, E., Weintraub, S. (1983). Boston Naming Test. Lea and Febiger.

Guy, W., 1976. Clinical Global Impressions. ECDEU Assessment Manual For Psychopharmacology, Revised (DHEW Publ No ADM 76-338). National Institute of Mental Health, Rockville, MD, pp. 218–222.

Haulcy, R., Glass, J., 2021. Classifying Alzheimer's disease using audio and text-based representations of speech. Front. Psychol. 11 (2021), 624137 https://doi.org/10.3389/fpsyg.2020.624137.
Helms, J.R., 2009. Mathematics for health sciences: a comprehensive approach. Cengage Learning.
Hernández-Domínguez, L., Ratte, S., Sierra-Martınez, G., et al., 2018. Computer based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's Dement 2018 10, 260–268. https://doi.org/10.1016/j.dadm.2018.02.004.
Higuchi, T., 1988. Approach to an irregular time series on the basis of the fractal theory. Physica D. 31 (1988), 277–283. https://doi.org/10.1016/0167-2789(88)90081-4.
Hirsh, K.W., Ellis, A.W., 1994. Age of acquisition and lexical processing in aphasia: a case study. Cogn. Neuropsychol. 11, 435–458. https://doi.org/10.1080/02643299408251981.
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. Psychol. Aging 33 (2), 195–218. https://doi.org/10.1037/pag0000236.
Kavé, G., Dassa, A, 2017. Severity of Alzheimer's disease and language features in picture descriptions. Aphasiology. https://doi.org/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. Front. Aging Neurosci. 11 (221) https://doi.org/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. Assess Dis Monit 1, 112–124, 2015.
Koo, J., Lee, J.H., Pyo, J., Jo, Y., Lee, K., 2020. Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. In: Proceedings of Interspeech 2020 (Shanghai), pp. 2217–2221. https://doi.org/10.21437/Interspeech.2020-3153.
Krein, L., Jeon, Y., Miller Amberber, A., Fethney, J., 2019. The assessment of language and communication in dementia: a synthesis of evidence. Am. J. Geriatr. Psychiatry 27 (4), 363–377. https://doi.org/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. Neurology, Psychiatry and Brain Res. 30, 137–143. https://doi.org/10.1016/j.npbr.2018.10.002. Volume2018, PagesISSN 0941-9500.
Liu, N., Yuan, Z., Chen, Y., Liu, C., Wang, L., 2023. Learning implicit sentiments in Alzheimer's disease recognition with contextual attention features. Front. Aging Neurosci. 15, 1122799 https://doi.org/10.3389/fnagi.2023.1122799.
Logsdon, R.G., Gibbons, L.E., McCurry, S.M., Teri, L., 1999. Quality of life in Alzheimer's disease: patient and caregiver reports. J. Ment. Health Aging 5, 21–32. https://doi.org/10.1111/j.1532-5415.1996.tb01405.x, 1999.
López-de-Ipiña, K., Alonso-Hernández, J.B., Solé-Casals, J., Travieso-González, C.M., Ezeiza, A., Faúndez-Zanuy, M., Calvo, P.M., Beitia, B., 2015. Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer's disease. Neurocomputing. 150, 392–401. https://doi.org/10.1016/j.neucom.2014.05.083, 2015.
Luz, S., Haider, F., Fuente, S.d.l., Fromm, D., MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. Proc. Interspeech 2020, 2172–2176. https://doi.org/10.21437/Interspeech.2020-2571.
Luz, S., la Fuente, S.D., Albert, P., 2018. A method for analysis of patient speech in dialogue for dementia detection. In: Kokkinakis, D (Ed.), 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)". ELRA, Paris.
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, pp. 3780–3784. https://doi.org/10.21437/Interspeech.2021-1220. August 30–September 3, 2021.
Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., MacWhinney, B, 2021b. Alzheimer's dementia recognition through spontaneous speech. Front. Comput. Sci. 3 (2021), 780169.
Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
Mauri, C., Ballarè, S., Goria, E., Cerruti, M., Suriano, F., 2019. KIParla corpus: a new resource for spoken Italian. In: Bernardi, R., Navigli, R., Semeraro, G. (Eds.), Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it.
McIntyre, M.C., 1994. Criterion-related and Construct Validation of the Disability Assessment For Dementia scale. Thesis submitted For an M.Sc. in Rehabilitation Science, School of Physical and Occupational Therapy. McGill University, Montreal, Canada, 1994.
Meghanani, A., Anoop, C.S., Ramakrishnan, A.G., 2021. Recognition of Alzheimer's dementia from the transcriptions of spontaneous speech using fastText and CNN Models. Front. Comput. Sci. 3, 624558 https://doi.org/10.3389/fcomp.2021.624558.
Mereu, D., Vietti, A., 2021. Dialogic ItAlian: The creation of a corpus of Italian spontaneous speech. Speech Comm. 130, 1–14. https://doi.org/10.1016/j.specom.2021.03.002.
Millington, T., Luz, S, 2021. Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. Front. Comput. Sci. 3 (2021), 649508 https://doi.org/10.3389/fcomp.2021.649508.
Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., Christensen, H., 2019. Dementia detection using automatic analysis of conversations. Computer Speech Lang 53, 65–79. https://doi.org/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. J. Psychiatr. Res. 43 (4 (2009)), 411–431. https://doi.org/10.1016/j.jpsychires.2008.04.014.
Mueller, K.D., Hermann, B., Mecollari, J., Turkstra, L.S., 2018. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. J. Clin. Exp. Neuropsychol. 40 (9), 917–939. https://doi.org/10.1080/13803395.2018.1446513.
Mura, T., Proust-Lima, C., Jacqmin-Gadda, H., Akbaraly, T.N., Touchon, J., Dubois, B., Berr, C., 2014. Measuring cognitive change in subjects with prodromal Alzheimer's disease. J. Neurol. Neurosurg. Psych. 85 (4), 363–370. https://doi.org/10.1136/jnnp-2013-305078.
Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al., 2005. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. (53), 695–699. https://doi.org/10.1029/WR017i002p00410.
Nevler, N., Ash, S., Irwin, D.J., Liberman, M., Grossman, M., 2019. Validated automatic speech biomarkers in primary progressive aphasia. In Ann. Clin. Transl. Neurol. 6 (1), 4–14.
Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., Zeman, D, 2016. Universal dependencies v1: a multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1659–1666.
Ostrand, R., Gunstad, J., 2021. Using automatic assessment of speech production to predict current and future cognitive function in older adults. J. Geriatr. Psychiatry Neurol. 34 (5), 357–369. https://doi.org/10.1177/0891988720933358, 2021.
Panisset, M., Roudier, M., Saxton, J., Boller, F., 1994. Severe impairment battery. A neuropsychological test for severely demented patients. Arch. Neurol. 51 (1), 41–45. https://doi.org/10.1001/archneur.1994.00540130067012, 1994 JanPMID: 8274108.
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. Rejuvenation. Res. 18 (5), 389–412. https://doi.org/10.1089/rej.2014.1637, 2015 Oct.
Petti, U., Baker, S., Korhonen, A., 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J. Am. Med. Inform. Assoc. 27 (11), 1784–1797. https://doi.org/10.1093/jamia/ocaa174.
Pope, C., Davis, B.H., 2011. Finding a balance: the carolinas conversation collection. Corpus. Linguist. Linguist. Theory. 7 (1), 143–161. https://doi.org/10.1515/cllt.2011.007.
Prins, R.S., Bastiaanse, R., 2004. Analysing the spontaneous speech of aphasic speakers. Aphasiology. 18, 1075–1091. https://doi.org/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. Association For Computational Linguistics (ACL) System Demonstrations, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9, 2019.
Reisberg, B., Ferris, S.H., de León, M.J., Crook, T., 1982. The global deterioration scale for assessment of primary degenerative dementia. Am. J. Psychiatry 139, 1136–1139. https://doi.org/10.1176/ajp.139.9.1136, 1982.
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. (Eds.), *Proceedings of the Workshop BioNLP 2007: Biological, Translational, and Clinical Language Processing*. ACL - Association for Computational Linguistics, pp. 1–8.
Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2081–2089. https://doi.org/10.1109/TASL.2011.2112351. https://ieeexplore.ieee.org/document/5710404.
Rosen, W., Mohs, R., Davis, K., 1984. A new rating scale for Alzheimer's disease. J. Psychiatric Res. 141 (11), 1356–1364. https://doi.org/10.1176/ajp.141.11.1356.
Schachter, P., Shopen, T., 2007. Parts-of-speech systems. In: Shopen, T (Ed.), Language Typology and Syntactic Description. Cambridge University Press, pp. 1–60.
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. Front. Comput. Sci. 3 (2021), 624659 https://doi.org/10.3389/fcomp.2021.624659.
Shopen, T., 2007. Language Typology and Syntactic Description. Cambridge University Press.
Sigona, F., Gili Fivela, B., Vigorelli, P., Bolioli, A., 2023. *A Computational Linguistic Analysis of Speech in Elderly Individuals with Cognitive Decline: The "Anchise-2022" Corpus*. Studi AISV, Offinicaventuno.
Solorio, T., Liu, Y., 2008. Using language models to identify language impairment in Spanish-English bilingual children. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 116–117.
Soria Lopez, J.A., González, H.M., Léger, G.C., 2019. Alzheimer's disease. Handb. Clin. Neurol. 167, 231–255. https://doi.org/10.1016/B978-0-12-804766-8.00013-3, 2019.
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, pp. 1031–1038.
Teng, E.L., Chui, H.C., 1987. The modified mini-mental state (3ms) examination. J. Clin. Psychiatry 48 (8), 314–318, 1987.
Themistocleous, C., Eckerström, M., Kokkinakis, D., 2018. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural in networks. Front. Neurol. 9, 1–10, 2018.
Themistocleous, C., Ficek, B., Webster, K., den Ouden, D.B., Hillis, A.E., Tsapkini, K, 2021. Automatic subtyping of individuals with primary progressive aphasia. J. Alzheimers. Dis. 79 (3), 1185–1194. https://doi.org/10.3233/JAD-201101.
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. Alzheimer's demen. Diagn. Assess. Dis. Monit. 10, 31–40. https://doi.org/10.1016/j.dadm.2017.08.005.
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. JAMa Intern. Med. 175 (9), 1450–1458. https://doi.org/10.1001/jamainternmed.2015.2152, 2015 Sep.
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: Whitaker, H.A. (Ed.), Neuropsychological Studies of Nonfocal Brain Damage: Dementia and Trauma. Springer, New York. https://doi.org/10.1007/978-1-4613-8751-0_4.
Vaughan, R.M., Coen, R.F., Kenny, R., Lawlor, B.A., 2018. Semantic and phonemic verbal fluency discrepancy in mild cognitive impairment: potential predictor of progression to Alzheimer's disease. J. Am. Geriatr. Soc. 66 (4), 755–759. https://doi.org/10.1111/jgs.15294.
Vigo, I., Coelho, L., Reis, S., 2022. Speech- and language-based classification of Alzheimer's disease: a systematic review. Bioengineering 9, 27. https://doi.org/10.3390/bioengineering9010027, 2022.
Vigorelli P. (ed.) (2004). La conversazione possibile con il malato Alzheimer. FrancoAngeli, Milano. ISBN: 9788846454553.
Vigorelli, P., 2010. The ABC group for caregivers of persons living with dementia: self-help based on the conversational and enabling approach. Nonpharmacol. Ther. Dement. 3, 271–286, 2010.
Vigorelli, P., 2011. L'approccio Capacitante. Come Prendersi Cura Degli Anziani Fragili e Delle Persone Malate Di Alzheimer. FrancoAngeli, Milano. ISBN: 9788856833133.
Vigorelli, P., 2021. Dialoghi Imperfetti. Per Una Comunicazione Felice Nella Vita Quotidiana e Nel Mondo Alzheimer. FrancoAngeli, Milano. ISBN: 9788835118008.
Vigorelli, P., 2024. The enabling approach, an Italian approach to persons living with dementia. Brain Sci. Neurosurg. 1.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R. (2018). GLUE: a multi-task benchmark and analysis platform for natural language understanding. 2018, arXiv preprint arXiv:1804.07461.
Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. (2019). Superglue: a stickier benchmark for general-purpose language understanding systems. 2019, arXiv preprint arXiv:1905.00537.
Weiner, J., Schultz, T., 2016. Detection of intra-personal development of cognitive impairment from conversational speech. In: Speech Communication; 12. ITG Symposium, pp. 1–5.
Yang, Q., Li, X., Ding, X., et al., 2022. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alz. *Re*.s Therapy 14, 186. https://doi.org/10.1186/s13195-022-01131-3 (2022).
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. Alz Res Therapy 109 (2021). https://doi.org/10.1186/s13195-021-00848-x, 13.