DOI: 10.17469/O2111AISV000012

FRANCESCO SIGONA, BARBARA GILI FIVELA, PIETRO VIGORELLI, ANDREA BOLIOLI

A computational linguistic analysis of speech in elderly individuals with cognitive decline: the "Anchise-2022"

In this article, we introduce the "Anchise-2022" corpus, the latest update of the "Anchise" corpus, with approximately 600 manually transcribed conversations, mainly between healthcare providers and elderly individuals with cognitive decline (dementia, Alzheimer's disease). We also analysed over 200 of these transcripts, to check if and to what extent we can correlate characteristics of such kind of speech production with the progression of dementia as measured by Mini-Mental State Examination (MMSE). To this purpose, we adopted computational linguistic analysis using Natural Language Processing (NLP) techniques, statistical-descriptive and inferential analysis. The results obtained allow us to assert, in line with existing literature, that certain parameters exhibit a degree of correlation with MMSE values. Additionally, it is feasible to utilize automatic classification techniques to infer the degree of linguistic/cognitive impairment.

Keywords: MMSE, NLP, Linguistic features, Ecological conversations, Enabling Approach.

1. *Introduction*

The importance of detecting language impairments has grown significantly in the diagnosis of various neurodegenerative diseases. These language deficits can manifest in different ways within neurodegenerative conditions. They may appear as a prominent symptom in the early stages, as seen in Primary Progressive Aphasia, or occur alongside other cognitive issues, as seen in Alzheimer's Disease (AD) (Boschi, Catricalà, Consonni, Chesi, Moro & Cappa, 2017). Additionally, individuals with Mild Cognitive Impairment (MCI) who later develop Dementia also exhibit language deficits (Karr, Graham, Hofer & Muniz-Terrera, 2018). Studies analysing connected speech in individuals with MCI related to AD and cognitively healthy individuals have highlighted significant differences, primarily related to reduced fluency and semantic impoverishment (Filiou, Bier, Slegers, Houzé, Belchior & Brambati, 2020). Furthermore, research comparing individuals with progressive AD and those who do not progress has found significant disparities in naming ability and semantic fluency (Kim, B.S., Kim, Y.B. & Kim, H., 2019). Language difficulties pose a substantial hurdle for dementia patients, particularly as the disease advances. In the early stages, language impairments typically involve difficulties in word retrieval, reduced verbal fluency, and compromised comprehension of written and spoken language (Banovic, Zunic & Sinanovic, 2018). As AD progresses, verbal fluency further declines, comprehension diminishes, and literal

and semantic paraphrases become more prevalent. In advanced stages, speech often becomes limited to repetitive echoing (echolalia) and verbal stereotypes (Soria Lopez, González & Léger, 2019). Given that analysing speech production closely mirrors real-life situations and daily functioning, it holds promise for assessing cognitive abilities and disease progression. However, traditional standardized psychometric instruments for evaluating linguistic abilities are time-consuming and prone to human bias. Recent studies have pointed out issues with standardization, normative data, criterion validity, and reliability of tools developed for non-neurodegenerative language and communication impairments (Krein, Jeon, Miller Amberber & Fethney, 2019). Additionally, these tools often overlook the perspective of individuals living with dementia (ILWD) regarding the impact of language impairments on their daily lives. Automatic analysis of speech and language (AASL) using Natural Language Processing (NLP) techniques has gained traction for assessing language proficiency, cognitive abilities, and even motor skills when audio signals are available. Numerous studies in the literature have focused on text-based AASL, correlating its features with psychometric scores and distinguishing between healthy and unhealthy subjects (Vigo, Coelho & Reis, 2022). However, there's a lack of consensus regarding the choice of tasks, features, and algorithms due to insufficient information (Gagliardi, 2023). Furthermore, most existing studies involve controlled experiments with pre-established tasks, which introduce some level of structure into speech output and result in semi-spontaneous speech (Prins, Bastiaanse, 2004). Research on spontaneous speech of ILWD in everyday interactive communication settings with high ecological validity has been limited. Conversations in ecological settings offer valuable insights, as they reflect a more genuine expression compared to controlled experiments. The Anchise Corpus, consisting of face-to-face conversations between ILWD in nursing homes and caregivers using the Enabling Approach, provides a unique dataset characterized by high ecological validity. It offers a substantial number of samples across various stages of cognitive decline, unlike smaller samples in controlled trials, resulting an essential resource for our investigation. This paper aims to evaluate the effectiveness of statistical analyses and classification approaches of NLP feature extracted by NLP techniques from transcribed dialogues in highly ecological contexts, in characterizing the oral language of speakers with dementia at various stages and identifying correlations between speech transcripts and patients' MMSE scores.

2. *Related work*

2.1 Elicitation tasks and speech resources

Adequate speech corpora for individuals with cognitive impairments are available for certain languages, such as English (e.g., TalkBank – DementiaBank, Becker, Boiler, Lopez, Saxton & Mcgonigle, 2004; for a comprehensive list of speech databases, see Vigo *et al.*, 2022). However, resources for the Italian language are notably scarce in comparison. Often, these resources involve a limited number

of participants and have not been compiled into published or readily accessible collections. The primary initiatives in this regard, which focus on analysing semi-spontaneous speech in specific Italian dialects, include the OPLON project (Beltrami, Gagliardi, Rossini Favretti, Ghidoni, Tamburini & Calzà, 2018; Calzà, Gagliardi, Rossini Favretti & Tamburini, 2021), the CIPP-ma corpus, and the CIPP-mci corpus (Dovetto, Guida, Pagliaro, Guardasci, Raggio, Sorrentino & Trillocco, 2022). These corpora have been assembled using well-established tasks. A frequently employed approach to collect speech samples involves having patients describe a scene, as demonstrated in the "Cookie Theft" picture (Goodglass, Kaplan & Weintraub, 1983). Traditionally, picture description has been considered more advantageous when compared to other discourse production tasks that have less structure, such as informal conversation or procedural tasks (Ulatowska, Allard, Donnell, Bristow, Haynes & Flower, 1988). Additionally, among various cognitively demanding tasks, recounting a story, whether it's a well-known tale like Pinocchio or Cinderella or a novel narrative presented through pictures (Toledo, Aluísio, Santos, Brucki, Trés, Oliveira *et al.*, 2017), requires the integration of characters and events within a temporal framework. Reading (with or without recall), narrating personal life experiences, dreams, and similar tasks are also commonly used. Petti, Baker & Korhonen (2020) and de la Fuente Garcia, Ritchie & Luz (2020) have provided comprehensive reviews of relevant studies in this area. Finally, in numerous instances, conversations between patients and doctors are documented during clinical assessments or cognitive evaluations. For example, in the study by Mirheidari, Blackburn, Walker, Reuber & Christensen (2019), neurologists were instructed to adhere to a predetermined set of questions explicitly designed to uncover common signs of impairments throughout the conversation. These questions included closed queries requiring long-term memory recall, compound questions, open-ended inquiries, and those related to memory concerns. Weiner and Schultz (2016) conducted an analysis of semi-standardized biographical interviews within the ILSE corpus (Weiner, Frankenberg, Telaar, Wendelstein, Schröder & Schultz, 2016). Espinoza-Cuadros, Garcia-Zamora, Torres-Boza, Ferrer-Riesgo, Montero-Benavides, Gonzalez-Moreira & Hernandez-Gomez (2014) recorded speech interactions from both clinicians and patients during structured interviews used in the administration of the Spanish version of the MMSE. Additionally, Luz, de la Fuente & Albert (2018) examined natural conversations taken from the Carolina Conversations Collection (Pope, Davis, 2011), which involved 21 ILWD and 17 individuals with no neuropsychological diseases (interviewers included gerontologists, linguistic students, or researchers).

2.2 Psychometric-clinical indices

There are various methods and scales available to assess the severity of cognitive impairments, dementia, and Alzheimer's disease (AD). The choice of method or scale depends on the specific context, whether it's research or clinical practice,

and the aspect being evaluated. For measuring overall outcomes, some commonly used scales include the Global Deterioration Scale (GDS) and the Clinical Global Impression of Change (CGI-C). To assess functional ability and quality of life, the Progressive Deterioration Scale (PDS) and the Disability Assessment for Dementia (DAD) are often employed. Cognitive ability is typically evaluated using various tools such as the Alzheimer's Disease Assessment Scale – cognitive subscale (ADAS-cog), the Severe Impairment Battery (SIB), the Montreal Cognitive Assessment (MoCA), and the Mini-Mental State Examination (MMSE, Folstein, M., Folstein, S & Mchugh, 1975). The Mini-Mental State Examination (MMSE) remains one of the most widely employed assessment tools today. It is a brief test, typically taking 10-15 minutes to complete, consisting of 30 items that assess various cognitive domains. These domains encompass temporal-spatial orientation, short-term memory, attention and calculation, the ability to recall memories, comprehension and language, nominal ability, as well as the ability to write sensibly and to replicate minimally complex designs. Studies have reported that the MMSE demonstrates internal consistency and exhibits short-term test-retest reliability in ILWD and long-term reliability in those with intact cognitive functioning. Additionally, it has proven to be sensitive to the severity of dementia in Alzheimer's disease (AD) patients. The total MMSE score is valuable for documenting cognitive changes over time, with AD patients typically experiencing an annual decline of 3 points on the MMSE (Bernard, Goldman, 2010). However, it's important to emphasize that the MMSE should not be solely relied upon for diagnosing dementia (Arevalo-Rodriguez, Smailagic, Roqué-Figuls, Ciapponi, Sanchez-Perez, Giannakou *et al.*, 2021). This caution arises from the potential influence of non-neurological factors on low scores, including factors like low education, language difficulties, and visual or auditory impairments. Other problematic aspects and limitations of the MMSE have been discussed in studies by Mitchell (2009), Tsoi (2015), and De Roeck (2019). In addition to its routine use in clinical practice, the Mini-Mental State Examination (MMSE) serves various other purposes. It is employed as an exclusion or inclusion criterion in clinical trials, primarily for screening cognitive impairment, and is included in the neuropsychological test batteries used in research studies. In Italy, the MMSE is a prerequisite for the accreditation of Nursing Homes under the National Health Service. It is typically administered to each patient within 30 days of admission and then repeated every 6 months according to the Individual Care Plan, except in cases of sudden changes that necessitate a new evaluation. Each Italian Region implements the national law with its specific Regional Council resolutions, such as DGR 7435/2001 and DGR 1765/2014 in Lombardy. Furthermore, the MMSE score is employed to categorize the severity of Alzheimer's disease (AD). For example, it is used to define a scale of severity in note n. 85 by the Italian Medicines Agency (AIFA)[1], which authorizes the prescription of AChE inhibitors (donepezil, rivastigmine, and galantamine) for mild and moderate AD, as well as memantine for moderate and severe AD. This note also outlines a staging

[1] https://www.aifa.gov.it/documents/20142/1728086/nota-85.pdf.

system based on MMSE scores: AD mild (MMSE 21-26), moderate (MMSE 10-20), moderately severe (MMSE 10-14), and severe (MMSE <10). This same scale is also adopted by the National Institute for Health and Clinical Excellence (NICE)[2] in England.

2.3 Analysis by means of linguistic features

Various methods can be employed to establish correlations between the linguistic production of speakers with cognitive impairments and clinical indexes like the MMSE. In several studies, linguistic features have been utilized for this purpose, resulting in correlation values that vary considerably. These variations depend on numerous factors, with the type of task or linguistic production playing a pivotal role. A comprehensive overview of these features and their presumed discriminatory power in dementia research can be found in de la Fuente Garcia *et al.* (2020: 1552) and Petti *et al.* (2020). Notably, the ADReSS Challenge at INTERSPEECH 2020, as described by Luz, Haider, de la Fuente Garcia, Fromm & Macwhinney (2020), defined two cognitive assessment tasks. The first task involved standard Alzheimer's speech classification, while the second task focused on neuropsychological score regression. In the latter task, participants were tasked with developing models to predict MMSE scores using audio recordings (and annotated transcriptions) obtained from participants who described the Cookie Theft picture (Luz *et al.*, 2021). A combination of acoustic and linguistic features directly extracted from audio recordings resulted in a baseline MMSE prediction root mean squared error (RMSE) of 5.28 (Luz *et al.*, 2020). A subset of participants reported and discussed their findings from the MMSE regression task, reporting a RMSE score between 4.54 (Balagopalan, Eyre, Robin, Rudzicz & Novikova, 2020) and 5.62 (Shah, Sawalha, Tasnim, Qi, Stroulia & Greiner, 2021). In Yeung, Iaboni, Rochon, *et al.*'s study (2021), a collection of features derived from Natural Language Processing (NLP) and Automated Speech Analysis (ASA) was extracted from 30 speech samples provided by individuals with Alzheimer's Disease (AD, MMSE: 15-20), those with MCI (MMSE: 23-26), and healthy controls (HCs, MMSE: 27-30), all having a minimum of 12 years of education. These same speech samples were evaluated by five clinicians, who rated them in terms of (1) word-finding difficulty, (2) incoherence, (3) preservation, and (4) errors in speech using a Likert scale (0=not present or normal finding; 1=mild; 2=moderate; 3=severe). The authors discovered statistically significant correlations between certain features and the clinical ratings. Similarly, Bueno-Cayo, Del Rio Carmona, Castell-Enguix, Iborra-Marmolejo, Murphy, Irigaray, Cervera & Moret-Tatay (2022) explored the relationship between MMSE scores and specific linguistic features in a cohort of 33 participants, consisting of 20 healthy controls and 13 individuals with Mild Cognitive Impairment (aged between 60 and 95, with basic or intermediate

[2] https://www.nice.org.uk/guidance/ta217/documents/alzheimers-disease-mild-to-moderate-donepezil-galantamine-rivastigmine-and-memantine-part-review-draft-scope2.

education). Their analysis revealed a positive correlation between MMSE scores and lexical density, which was measured as the ratio of semantic content words per 100 words in a sentence (Pearson's correlation = 0.488, $p<0.01$). Additionally, MMSE scores exhibited a negative correlation with age. However, no statistically significant correlation (α =0.05) was observed between MMSE scores and speech length, word frequency, or the number of tokens related to time, place, and action. In a multivariate linear regression analysis of MMSE scores, both lexical density and speech length emerged as significant predictors. The speech elicitation process involved a single open-ended question: "what are your plans for today and what are your plans for tomorrow," without any constraints on response time. However, in the study by Kavé, Dassa (2017), they observed negative correlations between MMSE scores and certain linguistic factors. Specifically, they found that MMSE scores had a negative correlation (Pearson) with the total number of words (-0.355, $p<0.5$) and with the mean word frequency (-0.339, $p<0.5$), indicating that greater dementia severity was associated with the production of more frequent words. On the other hand, MMSE scores were positively correlated with the types-to-token ratio (TTR) (0.572, $p<0.01$). These findings were based on a sample of 35 individuals with Alzheimer's disease (23 females, 12 males; age: 65-91; MMSE: 3-25; education: 8-16) who participated in a Cookie Theft picture description task. No statistically significant correlations were observed between MMSE scores and the content-word ratio, pronouns ratio, percentages of verbs, prepositions, and subordination markers. In contrast, Ostrand, Gunstad (2021) reported different results in their investigation. In the picture description task, they found positive correlations between the use of definite articles, determiners, and nouns as a percentage of the total word count and the Modified-Mini-Mental State Examination (3MS, Teng, Chui, 1987) scores. Additionally, they observed a negative correlation between the average lexical density and 3MS scores. Interestingly, in an expository speech task performed by the same individuals, only the average lexical density was found to be correlated with 3MS scores. The examples discussed here provide a glimpse into a complex landscape where different linguistic features can exhibit varying degrees of correlation with clinical indexes. These correlations are contingent upon several factors, with the type of task or linguistic production emerging as a pivotal determinant. For instance, in the study by Ostrand, Gunstad (2021), only lexical density showed a consistent correlation with 3MS scores across different tasks. Notably, this correlation was negative, contrasting with the positive correlation found in the study by Bueno-Cayo *et al.* (2022). It's important to highlight that, thus far, no strong correlation has been consistently observed between cognitive decline and linguistic and communication impairments (LCIs). Studies that have attempted to correlate linguistic characteristics with MMSE scores generally do not report a strong correlation. Some noteworthy exceptions do exist, primarily in works that focus on subjects with Alzheimer's disease (AD), such as the study by Fritsch, Wankerl & Nöth (2019), as opposed to more generalized cognitive impairment. In many studies aimed at identifying AD

or MCI, the approach involves categorization tasks and statistical comparisons of linguistic features between groups of unhealthy individuals (either AD or MCI) and matched healthy controls (HCs), as demonstrated in the work by Beltrami *et al.* (2018). Even when datasets encompass three or more diverse cohorts, most studies tend to limit their analyses to pairwise comparisons, typically contrasting dementia with healthy aging (Gagliardi, 2023). However, when considering the practical application of these findings in clinical contexts, the prevailing approach of pairwise comparisons may not be suitable for developing effective tools for real-life situations. This focus on pairwise comparisons often overlooks the intricate nature of assessing cognitive frailty within geriatric settings (Panza, Solfrizzi, Barulli, Santamato, Seripa, Pilotto & Logroscino, 2015). Some studies have exclusively considered or included comparisons between different "grades" of cognitive impairments, which aligns more closely with the approach presented in this work. For example, König, Satt, Sorin, Hoory, Toledo-Ronen & Derreumaux (2015) examined Alzheimer's disease (AD), MCI, and healthy controls (HCs) using vocal tasks like counting backward, sentence repetition, picture description, and verbal fluency. They identified significant differences ($p<0.05$) between AD and MCI only in the neuropsychiatric inventory (Cummings, Mega, Gray, Rosemberg-Thompson & Gornbein 1994). Only a limited number of research papers have delved into the detailed classification of dementia, with particular emphasis on subtyping frontotemporal degeneration (e.g., Fraser, Meltzer, Graham, Leonard, Hirst, Black & Rochon, 2014; Garrard, Rentoumi, Gesierich, Miller & Gorno-Tempini, 2014; Nevler, Ash, Irwin, Liberman & Grossman, 2019; Themistocleous, Eckerström & Kokkinakis, 2018, 2021; Cho, Nevler, Shellikeri, Ash, Liberman & Grossman, 2020). Regarding the classification techniques employed, numerous contributions have explored various machine learning algorithms, such as Support Vector Machines, Random Forests, and Decision Trees. These algorithms are used to automatically distinguish between healthy controls (HC) and individuals with MCI based on features extracted from acoustic and manually or automatically transcribed speech. These features encompass lexical, syntactic, semantic, and readability indexes, and researchers like Calzà *et al.* (2021) and Gagliardi, Tamburini (2021, 2022) have contributed to this area.

3. *Materials. The Anchise 2022 Corpus*

The Anchise Corpus, a valuable resource for research and training related to ILWD, has been continuously compiled since 2007 by the Anchise Group. This group comprises experts dedicated to the research, training, and care of ILWD. The corpus is founded on the principles of the "Enabling Approach," which was developed by one of the authors and collaborators back in 2004 (Vigorelli, 2004) and has evolved over subsequent years (Vigorelli, 2010, 2011; Lanzoni, Fabbo, Basso, Pedrazzini, Bortolomiol, Jones & Cauli, 2018; Vigorelli, 2021). The Enabling Approach is a non-pharmacological therapy for dementia that

aims to facilitate the speech of patients, even in the presence of memory and language disorders, with the goal of promoting a positive coexistence between individuals. To implement this approach, ILWD engage in conversations with trained operators from diverse backgrounds, including healthcare workers, educators, nurses, speech therapists, doctors, and psychologists, all of whom have undergone specific training in the Enabling Approach. According to the protocol, after initial greetings and obtaining informed consent, the operator initiates the conversation in a manner designed to encourage the patient to speak. This may involve phrases like "Could you tell me something about how you are going to spend your day?" or "What do you do during the day?" The operator then mostly remains silent, waiting for the elderly person's verbal production. When there is an interruption, the operator takes the floor and continues the conversation based on what the patient has said. Importantly, the operator does not lead the conversation but instead follows the patient's lead in their speech and world. The operator's role is not to diagnose or gather information but to participate in the conversation in a way that promotes the well-being of both speakers. Key techniques employed include refraining from asking questions, avoiding interruptions, refraining from corrections, echoing the patient's statements, and returning to the narrative motif. Other techniques involve listening attentively even when the speech is pathological and incomprehensible, respecting slowness and pauses, acknowledging and reflecting emotions, responding to questions, and communicating through gestures and tone of voice (Vigorelli, 2004, 2011). Conversations typically last 5-10 minutes, with no fixed duration, and conclude when signs of patient fatigue or intolerance become evident. It is important to note that since operators receive instructions in a training context, their adherence to the prompts may vary, unlike in a strictly controlled experimental setting. Since 2007, healthcare professionals affiliated with the Anchise Group have recorded, transcribed, and annotated these conversations, involving elderly individuals with cognitive impairment, primarily residing in Italian Nursing Homes. Specifically, the participants are Italian speakers with noticeable cognitive deficits, with inclusion criteria based on MMSE scores ranging from 0 to 28. Exclusion criteria encompass psychological or behavioural disturbances that hinder conversation, such as drowsiness, wandering, marked oppositional or aggressive behaviour, pronounced psychomotor agitation, severe dysarthria, and severe hypoacusia. The recordings are made with the prior informed consent of the relevant family member, the manager of the facility and, as far as possible, the elderly person participating in the conversation. The instrument is usually the operator's personal cell phone. The transcription is done by the conversant. A faithful transcription is required, including truncated words, neologisms, repetitions. The context annotations are freely indicated by the operator, someone does more, someone else does less. All transcriptions are incorporated into the Anchise Corpus and are intended for use in training sessions for Nursing Home operators, not for research purposes, without selecting them based on

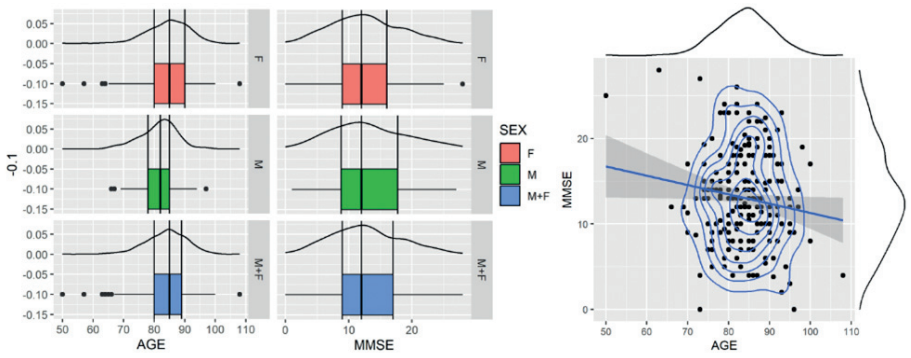
adherence to a specific style. For the same reason, the fidelity of the transcription is entrusted to the person sending the text and is not controlled. The corpus is continuously updated through the efforts of the Anchise Group. A prior version, the "Corpus Anchise 320," was released in 2020, comprising 320 conversations (Bolioli, Benvenuti, Mazzei & Vigorelli, 2020; Benvenuti, Bolioli, Bosca, Mazzei & Vigorelli, 2021). The Anchise 2022 Corpus extends this collection with 417 conversations. The significance of studying the Anchise Corpus lies in its representation of an Italian-language corpus collected in a natural environment. It serves as a valuable and rich source of information regarding the speech of individuals affected by dementia. The corpus reflects the diverse landscape found across Italy, including different regions (North, Center, and South), Nursing Homes of varying sizes, and patients of both sexes, with mild, moderate, and severe cognitive deficits. Patient classification is solely based on MMSE scores. Unlike prospective, controlled studies conducted under standardized conditions, this corpus was not originally created for research purposes but was intended for personnel training. Each dialogue or conversation within the corpus is organized into speech turns, with a new turn initiated each time there is a change in the speaker, continuing until the speaker talks without significant interruption. It is assumed that each conversation involves a different patient. Out of the 417 conversations, two were conducted in Spanish and English and were therefore excluded from the analysis. Some of the conversations include additional patient information, such as their gender, age, and corresponding MMSE score. For the experiments conducted, records were specifically selected based on patients with MMSE scores ranging from 0 to 26 (as detailed in § 4.1). These records were also required to include age information, providing more control over the subset of the corpus used in the experiments. This resulted in a total of 216 conversations. Table 1 provides initial insights into both the entire dataset and the filtered dataset. The interviewed patients have an average age of approximately 84 years, with a similar distribution between the entire corpus and the filtered records. The overall corpus is predominantly composed of female speakers, accounting for nearly 80 % of all speakers, and a similar gender ratio is observed in the filtered corpus. The conversations consist of an average of 65 turns, with a nearly even split between patient and operator turns. On average, turns contain over 14 tokens, although this figure significantly varies when considering patient turns (which tend to be longer, averaging around 20 tokens per turn) and operator turns (which are shorter, averaging about 9 tokens per turn). It is important to note that there is a high standard deviation, which is comparable to, if not greater than, the average itself, indicating substantial variation in turn length.

Table 1 - *Summary description of the Anchise 2022 Corpus composition with respect to subjects' type (Patients, Operators), patients' sex, speech turns, availability of MMSE score and age data*

	Corpus	*Filtered Corpus*
N. of conversations	417	216
Average patient age (stdev)	83.87 (7.09)	84.01 (6.93)
Male patients (percentage)	87 (21 %)	49 (23 %)
Female patients (percentage)	330 (79 %)	167 (77 %)
Average patient MMSE (stdev)	12.69 (5.60)	12.92 (5.46)
Average n. of turns per conversation	65.3	64.90
Average n. of patient turns per conversation	32.3	32.04
Average n. of operator turns per conversation	33	32.86
Average turn length (stdev)	14.25 (10.12)	14.83 (12.05)
Average turn length Patient (stdev)	19.60 (19.43)	20.31 (23.31)
Average turn length Operator (stdev)	8.85 (3.58)	9.24 (3.69)
Average conversation length (stdev)	831.68 (632.60)	843.50 (535.21)
Average conversation length – Patient (stdev)	543.76 (456.02)	545.82 (395.64)
Average conversation length – Operator (stdev)	287.92 (236.32)	297.67 (214.42)

Conversations in the filtered corpus are slightly longer than those in the broader corpus (averaging 843 tokens compared to 831 tokens). Again, this average length is accompanied by a notable standard deviation, suggesting consistent variations in length. Patient conversations are nearly twice as long as operator conversations. As depicted in Figure 1, their joint distribution of MMSE and age follows an approximately Gaussian distribution (Henze-Zirkler test = 0.86, p-value=0.15). There is a weak negative correlation between MMSE scores and age (β=-0.11, 95 % CI=[-0.21, -0.00426], t(217)=-2.05, p=0.041).

Figure 1 - *Distribution of the sessions with respect to Sex, Age, and MMSE scores. Left: boxplots with kernel density estimation (k.d.e.) for Age and MMSE by female, male and both. Right: scatterplot (dots), bivariate contour lines, linear trend and marginal k.d.e., for Age and MMSE (both male and female)*



4. *Method*

4.1 MMSE clustering and Evaluation Measures

In the original work, Folstein *et al.* (1975) reported that the MMSE should be scored from 0 to 30, with a score of 24 or greater as "normal" and with a score less than 20 "likely dementia." Afterwards, different classification scales (i.e. cut-off values) based on the MMSE score have been proposed or used in the literature. Some of these are illustrated as examples in Table 2 (see also Monroe, Carter, 2012, for further considerations). In this work only one severity scale was used, and the choice fell on the one recommended by AIFA, due to the national importance of the Agency, also supported by the adoption of the same scale abroad. However, the study could also be replicated for other scales. In the following, for sake of simplicity, the "AIFA85" term will be used to indicate the stages of cognitive impairments as depicted in the last row (*) of Table 2. The classification experiments considered only SEVERE, MODERATE and MILD stages, since the category of MMSE 27-30, has only 2 individuals. As can be seen by comparing (5) and (*) rows of Table 2, the "moderate-severe AD" (10-14) and "moderate AD" (15-20) grades were merged into a single "MODERATE" (10-20) grade. The reason for this merger lies primarily in the need to avoid complex classifications into four classes. Secondly, this grouping is also provided for by note 85 of AIFA, in the prescription of medications for patients with AD (memantine in severe patients with MMSE<=10; denoepezil, rivastigmine, galantamine, memantine, in moderate patients, with MMSE in 10 – 20, and donepezil, rivastigmine, galantamine in mild patients, with MMSE in 21 and 26).

Table 2 - *Some of the MMSE-based classification scales used in the literature. (1) De Stefano, Di Giovanni, Kulamarva, Di Fonzo, Massaro, Contini, Dispenza & Cazzato, 2021; (2) Tombaugh, McIntyre, 1992; (3): Chopra, Cavalieri & Libon, 2007; (4): Beltrami et al., 2018; (5): note n. 85 by the Italian Medicines Agency (AIFA), and National Institute for Health and Clinical Excellence (NICE), England. (*) = staging adopted in the present work, with number of individuals in parentheses. (mod.=moderate)*

<table>
<tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td><td>13</td><td>14</td><td>15</td><td>16</td><td>17</td><td>18</td><td>19</td><td>20</td><td>21</td><td>22</td><td>23</td><td>24</td><td>25</td><td>26</td><td>27</td><td>28</td><td>29</td><td>30</td><td></td></tr>
<tr><td colspan="10">Severe cogn. decline</td><td colspan="11">Mod. cogn. decline</td><td colspan="5">MCI</td><td colspan="5">Normal</td><td>(1)</td></tr>
<tr><td colspan="18">Severe cogn. impairment</td><td colspan="7">MCI</td><td colspan="5">Normal</td><td>(2)</td></tr>
<tr><td colspan="13">Severe impairment</td><td colspan="5">Mod. impair.</td><td colspan="7">Mild impair.</td><td colspan="3">Border-line</td><td colspan="3">Normal</td><td>(3)</td></tr>
<tr><td colspan="18"></td><td colspan="7">MCI or early dementia</td><td colspan="5">Normal</td><td>(4)</td></tr>
<tr><td colspan="9">Severe AD</td><td colspan="6">Mod.-severe AD</td><td colspan="8">Mod. AD</td><td colspan="6">Mild AD</td><td colspan="3"></td><td>(5)</td></tr>
<tr><td colspan="9">Severe (60)</td><td colspan="14">Mod. (137)</td><td colspan="6">Mild (19)</td><td colspan="3">Other (2)</td><td>(*)</td></tr>
</table>

Identifying correlations with MMSE scores involves developing evaluation tools capable of taking a transcript of a specific conversation as input and predicting the corresponding MMSE score for the patient participating in that conversation. Currently, we calculate

the Pearson and Spearman correlation coefficients (Benesty, Chen, Huang & Cohen, 2009) for this purpose. To evaluate the accuracy of the classification experiments, we used scores popular for Information Retrieval tasks, such as macro Precision, Recall and F1. In the multiclass classification task, for each class *c* we define:

- a *precision* score, as the number of samples of the class *c* that are correctly classified, divided by the total number of samples that are classified as *c*. The precision is the fraction of correct answers provided by the system.
- a *recall* score, as the number of samples of the class *c* that are correctly classified, divided by the number of input samples of the same class *c*. In other words, recall is the proportion of samples of class *c* that are correctly classified.
- a *F1* score, as the harmonic mean between precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The macro Precision is defined as the arithmetic mean of the precision across all the classes, and so macro Recall and macro F1.

4.2 Linguistic features

Leveraging the NLP tools of the "Stanza" software (Qi, Zhang, Y., Zhang, Y., Bolton & Manning, 2020), we extracted a range of morphosyntactic information and dependency relationships among the parts of speech from every turn within each patient session in the corpus (see Figure 2), and from these, also the linguistic features at the lexical, syntactic, and semantic levels shown in Table 3, for each session in the Corpus.

Figure 2 - *Top: a dependency tree among words of the sentence: Ecco... Io sono stato contento di essere qua con voi qua oggi ... ('well... I was happy to be here with you here today ...'). Bottom: a tabular representation of the parse tree, equipped with part-of-speech analysis obtained through the Stanza parser for the same sentence, in the CONLL-U format (Nivre, De Marneffe, Ginter, et al., 2016)*

id	form	lemma	upos	xpos	head	deprel	feats
1	Ecco	ecco	INTJ	I	6	discourse	
2	...	...	PUNCT	FS	1	punct	
3	Io	io	PRON	PE	6	nsubj	Number=Sing|Person=1|PronType=Prs
4	sono	essere	AUX	VA	6	aux	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
5	stato	essere	AUX	V	6	cop	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
6	contento	contento	ADJ	A	0	root	Gender=Masc|Number=Sing
7	di	di	ADP	E	9	mark	
8	essere	essere	AUX	V	9	cop	VerbForm=Inf
9	qua	qua	ADV	B	6	advmod	
10	,	,	PUNCT	FF	6	punct	
11	con	con	ADP	E	12	case	
12	voi	voi	PRON	PE	6	obl	Number=Plur|Person=2|PronType=Prs
13	qua	qua	ADV	B	14	advmod	
14	oggi	oggi	ADV	B	12	advmod	
15	...	...	PUNCT	FS	6	punct	

Table 3 - *Linguistic features*

Feature name (Lexical)	*Description*
content_density	the ratio between open-class words and open + closed class words (Roark, Mitchell, Hosom, Hollingshead & Kaye, 2011)
open_to_close_class	the ratio of open-class words to closed-class words
tokens_to_turns	the ratio between tokens and speech turns
num_words	the number of words
num_interrog	the number of interrogative sentences
num_ellipsis	the number of ellipsis «...»
word_length	the average number of characters of the words[3]
(Semantic & Lexical richness)	
num_hapax_lego	the number of words that appear only once in the session
num_types	the number of word types (i.e. different words)
ttr	types to token ratio
brunet indices	Brunet indexes. A measure of lexical diversity, used in stylometric analyses of text, often claimed to be independent of text length: $$BI = N^{V^{(-a)}}$$ where N is the text length, V is the number of different words, and −a is a scaling constant in a {0.134, 0.172, 0.185}
honore	Honoré statistics. An index of vocabulary richness, based on the idea that texts with richer vocabulary have a higher proportion of words that are hapax legomena. $$R = 100 \times \frac{\log(N)}{1 - \frac{v_1}{V}}$$ where N is the number of words v1 is the number of hapax, and V is the number of types
(Lexical / syntactic)	
<pos>_perc	Rate of: Adjectives (ADJ), Adpositions (ADP), Adverbs (ADV), Auxiliary (AUX), Coordinate conjunctions (CCONJ), Subordinate conjunctions (SCONJ), Nouns (NOUN), Pronouns (PRON), Interjections (INTJ), Proper nouns (PROPN), Verbs (VERB) E.g. adj_perc is the rate of Adjectives (ADJ) per 100 words.
sconj_to_cconj	ratio between subordinate and coordinate conjunctions (sentences)
nouns_to_verbs	nouns to verbs ratio. Also referred to as "reference rate to reality" (Vigorelli, 2004)
pronouns_to_nouns	the ratio between pronouns and nouns
x_perc	Rate of words that could match any other category (X) per 100 words.

[3] Even if there is no perfect correspondence between phone and characters, we considered this metric as it appears in other articles, such as in Orimaye, Wong, J.S., Golden, Wong, C.P. & Soyiri, 2017 and given that the Anchise corpus lacks audio tracks, which would be needed to count the number of the syllables and phones which have actually been realized.

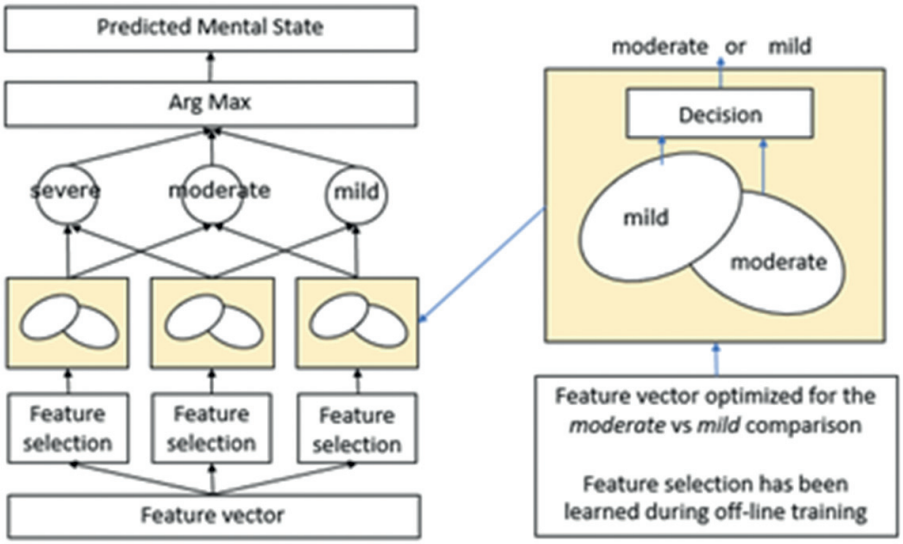
(Syntactic)	
dependency distance	Dependency distance (Roark, Mitchell & Hollingshead, 2007; Roark *et al.*, 2011). Mean, standard deviation and maximum values have been considered across each sentence.
max_depth	Maximum structure depth (mean, standard dev., maximum across each turn)
szmrecsanyi	A syntactic complexity measure by Szmrecsanyi, 2004: $\frac{2 * conj + 2 * pron + nouns + verbs}{words}$ Since subordinators and pronouns are considered as the clearest indicators of increased embeddedness (and thus of high complexity), these features have a higher weight than verbal forms and noun phrases.
(Verbal analysis)	
va_fin_rate, va_inf_rate, va_part_rate, va_ger_rate	Rates of finite, infinitive, participle, gerund verbal forms
reduced_sentences_count	n. of participle and gerund verbal forms
reduced_sentences_to_verbs	Ratio of participles and gerunds to total verbs
va_ind_rate, va_subj_rate, va_imp_rate, va_cond_rate	Rates of indicative, subjunctive(and conjunctive), imperative and conditional verbal moods
va_pres_rate, va_past_rate, va_imperf_rate, va_future_rate	Rates of present, past, imperfect and future tense
va_p1_rate	Rate of 1st person verbs

4.3 Correlation and classification between Linguistic Features and MMSE Scores

The features will be subject to analysis through linear regression, as well as Pearson and Spearman correlations, in relation to the MMSE scores.

Additionally, these same features will be employed for conducting statistical comparisons among different clusters of MMSE scores, categorized based on the AIFA85 stages. Due to variations in sample size across stages and the relatively small sample size in some cases, we have opted for the non-parametric two-sample Kolmogorov-Smirnov (KS) test. This test assesses the distance between the empirical distributions of individual features in both classes. We have established a significance level of α=0.05 as the threshold to determine statistical significance. Furthermore, we will conduct a subject classification task, which involves predicting the category of a session based on its linguistic features. This task is accomplished using the tool illustrated in Figure 3.

Figure 3 - *Scheme of the multiclass classifier used for the linguistic features classification experiment*



This tool consists of three binary classifiers, each designed to predict between two classes. The ultimate prediction for an input session is determined by the most frequently predicted output class. Each binary classifier is trained independently from the others, using only the sessions that belong to the two relevant classes. A leave-one-out cross-validation (LOOCV) approach is employed, where each session is taken out from the dataset and utilized for prediction, while the remaining sessions serve as the training dataset. We have tested two types of binary classifiers. The first assumes that the statistical distribution of the linguistic features for each of the two classes follows a Multivariate Normal distribution. During testing, it selects the class with the highest likelihood given the input session. The second classifier is based on Linear Discriminant Analysis (LDA), implemented using the Matlab® fitcd iscr function. During the training process, feature selection algorithms are applied to optimize performance metrics and identify a set of features that are especially responsive to the classification task. In this regard, we have explored Forward Feature Selection (FFsel) and Forward-Backward Feature Selection (FBsel). It is worth noting that the feature combination obtained through these methods may not necessarily represent the optimal choice, as there could be other combinations that yield equal or superior performance.

5. *Results*

5.1 Correlation between linguistic features and MMSE

Overall, the correlations between MMSE scores and individual linguistic features were relatively weak. However, some of these correlations did reach statistical significance ($p<0.05$). Table 4 displays these significant correlations, with the highest value being 0.19, representing the Pearson's correlation coefficient between MMSE scores and the frequency of verbs in the subjunctive mood. Regarding positive correlations, there are significant associations between MMSE scores and several linguistic features. Both Pearson and Spearman coefficients indicate a positive correlation between MMSE scores and the count of hapax legomena (i.e., words spoken only once in a session) as well as the rate of verbs in the subjunctive mood. Additionally, the rate of nouns, the rate of verbs in the past tense, the count of reduced sentences (i.e., the number of participle and gerund verbal forms), and the mean value of dependency distance are significantly correlated, though with only one of the correlation coefficients. Conversely, both Pearson and Spearman coefficients reveal significant negative correlations between MMSE scores and the rates of interjections, finite verbs, and the rate of verbs in the present tense. There is also a negative correlation, according to the Pearson coefficient, with the adjectives rate. Several features, including content density, showed no significant correlation with MMSE scores. Overall, while correlations between linguistic features and MMSE scores are weak, our findings align with existing literature.

Table 4 - *Linear fit (beta, value of the statistic, confidence interval, p-value and R-squared) and correlation values (Pearson, Spearman) between each single linguistic feature and MMSE scores. Only statistically significant results are reported (*p < 0.05, **p < 0.01, ***p < 0.001). Reduced sentence count is defined as the number of participle and gerund verbal forms. The label "(mean value of dependency distance" refers to the mean value of dependency distance. (r.o. = rate of)*

features	*beta*	*t*	*R2*	*Pearson*		*Spearman*	
				r	*p-value*	*ρ*	*p-value*
hapax count	0.0132	t(214)=2.250	0.023	0.152	0.025 *	0.145	0.033 *
r.o. adjectives	-0.4252	t(208)=-2.368	0.026	-0.162	0.019 *	-0.076	0.270
r.o. interj.	-0.1340	t(213)=-2.422	0.027	-0.164	0.016 *	-0.157	0.021 *
r.o. nouns	0.2124	t(212)=1.885	0.016	0.128	0.061 .	0.167	0.014 *
r.o. finite verbs	-0.0770	t(214)=-2.693	0.033	-0.181	0.008 **	-0.160	0.019 *
r.o. verbs in subjunctive mood	0.2622	t(214)=2.834	0.036	0.190	0.005 **	0.191	0.005 **
r.o.verbs in present tense	-0.0457	t(214)=-2.264	0.023	-0.153	0.025 *	-0.142	0.037 *
r.o. verbs in the past tense	0.0650	t(214)=2.054	0.019	0.139	0.041 *	0.120	0.078 .
n. reduced sentences	0.0374	t(214)=1.448	0.010	0.098	0.149	0.156	0.022 *
Dep dist mean avg	1.8419	t(214)=2.100	0.020	0.142	0.037 *	0.123	0.070 .

For example, we observed a positive correlation between the count of hapax legomena (words spoken only once) and MMSE scores, which is consistent with previous research indicating that individuals with greater cognitive impairment tend to have a more limited vocabulary. Additionally, our discovery of a positively correlated rate of nouns is in line with the concept of anomia, the initial linguistic disorder that often appears in individuals with cognitive impairments, particularly those with Alzheimer's disease (AD). Furthermore, the negative correlation we found between the rate of interjections and MMSE scores may suggest an increase in the level of uncertainty in speech production as MMSE scores decrease. These findings are consistent with previous research in the field. With respect to the mood and tense of verbal forms, as MMSE scores decline, we observe a trend toward simplification in the use of verbs. This trend aligns with real-world observations, indicating that individuals tend to employ more finite forms while reducing the use of participles, gerunds, and subjunctive moods. Moreover, there is a shift in the use of tenses, with a flattening effect towards the present tense at the expense of the past tense, as part of an overall simplification process. It's worth noting that the utilization of the future tense remains quite limited in these contexts. Several other features did not exhibit statistically significant correlations with MMSE scores. However, it's important to consider that certain speech-related or task-specific features may play a role. For instance, content density appears to be highly influenced by the nature of the speech task. To illustrate, a study by Beltrami, Calzà, Gagliardi, Ghidoni, Marcello, Rossini Favretti & Tamburini (2016) revealed that content density had statistical significance when comparing individuals with Mild Cognitive Impairment (MCI) to healthy controls in the context of a picture description task but not in a narrative task.

5.2 Classification

A statistical analysis was done on the distribution of each feature across the AIFA85 stages independently, utilizing the omnibus Kruskal-Wallis (KW) test and the Kolmogorov-Smirnov (KS) test. The KS test was applied even when the omnibus test did not yield significant results. Results are displayed in Table 5.

Table 5 - *Statistically significant differences between the distribution of the linguistic features in the AIFA85 stages. The statistics values and p-values of the Kruskal-Wallis (KW) and pairwise Kolmogorov-Smirnov tests are reported for each comparison (* p < 0.05, ** p < 0.01, *** p < 0.001). Bold values are significant after Bonferroni correction (p<0.05/3). (dep_dist. mean.avg = mean value of dependency distance)*

		SEVERE vs MODERATE d.o.f. (60, 137)		*SEVERE vs MILD d.o.f. (60,19)*		*MODERATE vs MILD d.o.f. (137,19)*	
	KW	*KS*	*p-value*	*KS*	*p-value*	*KS*	*p-value*
words count		0.167	0.165	0.384	0.019 *	0.257	0.182
hapax count	6.29 *	0.180	0.106	*0.428*	*0.006* **	0.334	0.035 *

		SEVERE vs MODERATE d.o.f. (60, 137)		*SEVERE vs MILD d.o.f. (60,19)*		*MODERATE vs MILD d.o.f. (137,19)*	
	KW	*KS*	*p-value*	*KS*	*p-value*	*KS*	*p-value*
average n. of words per turn		0.135	0.392	0.348	0.045 *	0.230	0.290
word-types count		0.130	0.419	*0.395*	*0.015* *	0.290	0.094 .
nouns to verbs ratio	6.32 *	*0.256*	*0.007* **	0.298	0.121	0.273	0.134
adpositions rate		0.118	0.550	0.361	0.034 *	0.260	0.171
nouns rate	8.25 *	*0.261*	0.005 **	0.279	0.170	0.192	0.499
rate of verbs in the subjunctive mood	11.3 **	*0.234*	*0.011* *	*0.439*	*0.002* **	0.264	0.135
dep_dist.mean.avg		0.147	0.294	0.359	0.036 *	0.258	0.179

In comparing the distribution of each single feature (independently of any other) in the 3 pairs formed by the 3 groups SEVERE, MODERATE and MILD, it is appropriate to adopt the Bonferroni correction. Therefore, results are presented and discussed with respect to a significance level α=0.05 as well a corrected α=0.05/3 = 0.017. The results indicate that four features are significant at the Kruskal-Wallis test, namely: the hapax count, the nouns-to-verbs ratio, the nouns rate and the rate of verbs in the subjunctive mood. Among these, three features, specifically the nouns to verbs ratio, nouns rate, and the rate of verbs in the subjunctive mood, display statistically different distributions when comparing SEVERE and MODERATE stages. Additionally, only one feature, the count of hapax legomena, exhibits distinct distribution between the MODERATE and MILD categories, but not after the Bonferroni correction. In contrast, eight features (three, after the Bonferroni correction) exhibit significant differences in distribution between the extreme categories, SEVERE and MILD. When comparing these results with the correlation analysis, it becomes evident that more features related to vocabulary richness exhibit statistical significance when distinguishing between the SEVERE and MILD stages. Specifically, the number of word-types (i.e., different words) and the count of hapax legomena, become significant (at different significance levels). This implies that the decline in vocabulary richness becomes more pronounced in the more severe stage. Notably, the significance of the hapax count, when shifting from MILD to MODERATE, is weaker compared to the comparison between the extreme categories. In the classification task, all available features were utilized, irrespective of whether they had been identified as significant or not. This approach aimed to ensure that no potential beneficial interactions between significant and non-significant features were overlooked. The most effective variant of the classifier discussed in § 4.3 employed multivariate normal (MVN) modelling of the feature distribution and employed Forward Backward selection (FBsel) as a feature selection strategy. The resulting evaluation metrics are presented in Table 6, which includes macro F1 scores, precision, and recall

scores for both the 3-classes and each individual 2-classes classification task. Most studies in the literature typically tackle a binary classification problem (e.g., AD vs. healthy or AD vs. MCI). Therefore, when comparing our results, we primarily focus on this aspect. In this context, our findings reveal that the highest accuracy is achieved in the classification of SEVERE vs. MILD. The F1 score, while not identical, is reasonably close to the results presented by Calzà *et al.* (2021), which also include some comparisons with other findings from the literature. However, it's essential to note that a direct comparison may not be entirely appropriate since the study by Calzà *et al.* involved the classification between MCI and healthy controls (F1=0.7045), and their feature set was more extensive, encompassing acoustic features as well.

Table 6 - *Summary results of the classification of the linguistic features. Macro F1, precision and recall*

Classification task	*F1*	*Precision*	*Recall*
SEVERE vs. MODERATE vs. MILD	0,533	0.562	0.537
SEVERE vs. MODERATE	0.630	0.627	0.636
SEVERE vs. MILD	0,779	0,785	0,773
MODERATE vs. MILD	0.596	0.595	0.691

6. *Discussion*

Analysing the transcribed dialogues obtained in ecologically relevant settings posed several challenges, especially when detailed sociolinguistic and diagnostic information was lacking. Nonetheless, despite its unique characteristics, both in terms of its objectives and data collection methods, the Anchise Corpus provided valuable qualitative and quantitative insights into speech patterns in individuals with dementia. Overall, the results of the analysis are in line with those reported in the literature. In summary, we did observe correlations between linguistic features in oral production and MMSE scores, although these correlations were generally weak. These findings align with the mixed results reported in the literature, as discussed in § 2.3, and may be attributed in part to the specific task and speech style we investigated, consistent with the outcomes in Beltrami *et al.* (2016). Nevertheless, the significant correlations we identified are in accordance with existing literature. For example, the positive correlation between the count of hapax legomena and MMSE scores supports findings in Hernandez-Dominguez, Ratte, Sierra-Martinez, *et al.* (2018). In essence, the significant correlations imply that individuals with more pronounced cognitive impairment tend to exhibit a reduced vocabulary and an increased level of uncertainty in their speech production, as evidenced by a higher frequency of interjections. Additionally, they appear to simplify their use of verbs by relying more on finite modes and present tense forms. Regarding the task of distinguishing between different stages of cognitive impairment based on MMSE

scores, the tools utilizing lexical and morphosyntactic features delivered promising accuracy results. For instance, in the binary classification task of distinguishing SEVERE from MILD cases, the F1 score approached 0.78. In the broader 3-class categorization task, the macro-averaged F1 score reached 0.53. These results stand up well to those reported in the literature, even when additional information and features, such as audio signals, were incorporated into the classification. Notably, the literature often focuses on distinguishing impaired individuals from healthy controls, as seen in the study by Calzà *et al.* (2021).

7. *Conclusions and Future Work*

Our research aimed to present the new extended release of the Anchise corpus, and to gain a deeper understanding of the speech patterns of ILWD engaged in natural conversations. We explored how linguistic features derived from NLP of the transcribed turn of speech could be correlated to the level of cognitive impairment, as assessed by the MMSE score. To achieve this goal, we analysed over 200 transcribed conversations from the Anchise 2022 Corpus. Our research revealed that there is generally a weak correlation between individual speech features and MMSE scores when examined individually. However, these features proved to be reasonably effective in distinguishing ILWD within MMSE-based clustering systems like NICE or AIFA85, particularly when differentiating between severe and mild cognitive impairments. This outcome aligns with the fact that our NLP experiments and MMSE tests focus on partially different aspects, as MMSE assessments are not specific to language or speech capabilities. This is the first experiment on a large corpus of spontaneous dialogues in Italian, but at the same time it is only a starting point. Indeed, the analysis of spontaneous speech by ILWD using the Anchise Corpus resources will be further expanded. In future works, we plan to integrate various analytical approaches, including a broader set of linguistic features, as well as the use of automatic emotion analysis and classification techniques employing Deep Language Models. This more comprehensive range of analyses is expected to provide a more detailed understanding of linguistic and communication deficits and offer more precise insights into the progression of dementia in relation to MMSE scores, surpassing the results obtained in this study.

Acknowledgements

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

Bibliography

Arevalo-Rodriguez, I., Smailagic, N., Roqué-Figuls, M., Ciapponi, A., Sanchez-Perez, E. & Giannakou, A. et al. (2021) Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). In *Cochrane Database of Systematic Reviews*, 2021(7), CD010783. doi: 10.1002/14651858.CD010783.pub3

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F. & Novikova, J. (2021). Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. In *Frontiers in aging neuroscience*, 13(2021): 635945. doi: 10.3389/fnagi.2021.635945

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

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J. & McGonigle, K.L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. In *Arch Neurol*, 1994; 51(6):585–94.

Beltrami, d., calzà, l., gagliardi, g., ghidoni, e., marcello n., rossini favretti, r. & tamburini, f. (2016). Automatic identification of Mild Cognitive Impairment through the analysis of Italian spontaneous speech productions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2086–2093, Portorož, Slovenia. European Language Resources Association (ELRA).

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F. & Calzà, L. (2018). Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? In *Front. Aging Neurosci*. 2018, 10, 369. doi: 10.3389/fnagi.2018.00369

Benesty, J., Chen, J., Huang, Y. & Cohen, I. (2009). Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-00296-0

Benvenuti, N., Bolioli, A., Bosca, A., Mazzei, A. & Vigorelli, P. (2021). The "Corpus Anchise 320" and the analysis of conversations between healthcare workers and people with dementia. In Dell'Orletta, F., Monti, J. & Tamburini, F. (a cura di), *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*: Bologna, Italy, March 1-3, 2021. Torino: Accademia University Press. doi:10.4000/books.aaccademia.8260, 51-57.

Bernard, B., Goldman, J.G. (2010). MMSE – Mini-Mental State Examination. In Kompoliti K., Verhagen Metman, L. (a cura di), *Encyclopedia of Movement Disorders*, Academic Press, 2010, 187-189, doi: 10.1016/B978-0-12-374105-9.00186-6

Bolioli, A., Benvenuti, N., Mazzei, A. & Vigorelli, P. (2020). Analisi linguistica computazionale del "Corpus Anchise" di dialoghi operatore-paziente. In *Atti del 65° Congresso Nazionale Società Italiana di Gerontologia e Geriatria SIGG 2020*, Italy, 2-4 December, 2020.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A. & Cappa, S.F. (2017). Connected speech in neurodegenerative language disorders: a review. In *Front Psychol*. 2017; 8:269. doi:10.3389/fpsyg.2017. 00269

Bueno-Cayo, A.M. del Rio Carmona, M., Castell-Enguix, R., Iborra-Marmolejo, I., Murphy, M., Irigaray, T.Q., Cervera, J.F. & Moret-Tatay,

C. (2022). Predicting Scores on the Mini-Mental State Examination (MMSE) from Spontaneous Speech, In *Speech. Behav. Sci.* 2022, 12, 339. doi: 10.3390/bs12090339

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

Cho, S., Nevler, N., Shellikeri, S., Ash, S., Liberman, M.Y. & Grossman, M. (2020). Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features. In *RaPID@LREC*.

Chopra, A., Cavalieri, T.A. & Libon D.J. (2007). Dementia screening tolls for the primary care physician. In *Clin Geriatr*. 2007;15(1):38–45.

Cummings, J.L., Mega, M.S., Gray, K., Rosemberg-Thompson, S. & Gornbein, T. (1994). The neuropsychiatric inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 1994; 44:2308–14. doi: 10.1212/wnl.44.12.2308

de la Fuente Garcia, S., Ritchie, C.W. & Luz, S. (2020). Artificial Intelligence, speech, and language processing approaches to monitoring Alzheimer's Disease: A systematic review. In *Journal of Alzheimer's Disease*, 78(4), 1547–1574. doi: 10.3233/JAD-200888

De Roeck, E.E., De Deyn, P.P., Dierckx, E. & Engelborghs, S. (2019). Brief cognitive screening instruments for early detection of Alzheimer's disease: a systematic review. In *Alz Res Therapy*, 11, 21(2019). doi: 10.1186/s13195-019-0474-3

De Stefano, A., Di Giovanni, P., Kulamarva, G., Di Fonzo, F., Massaro, T., Contini, A., Dispenza, F. & Cazzato, C. (2021). Changes in Speech Range Profile Are Associated with Cognitive Impairment. In *Dement Neurocogn Disord*. 2021 Oct;20(4):89-98. doi: 10.12779/dnd.2021.20.4.89

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

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

Filiou, R.P., Bier, N., Slegers, A., Houzé, B., Belchior, P. & Brambati, S.M. (2020). Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. In *Aphasiology*, 34(6), 723–755. doi: 10.1080/02687038.2019.1608502

Folstein, M., Folstein, S. & McHugh, P. (1975). "Mini-mental state"—A practical method for grading the cognitive state of patients for the clinician. In *J. Psychiatric Res.*, vol. 12, pp. 189–198, 1975. doi: 10.1016/0022-3956(75)90026-6

Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E. & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. In *Cortex*, 55, 43–60.

Fritsch, J., Wankerl, S. & Nöth, E. (2019). Automatic diagnosis of Alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing*. IEEE; 2019, 5841–5. doi: 10.1109/ICASSP.2019.8682690

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

Gagliardi, G., Tamburini, F. (2021). Linguistic biomarkers for the detection of Mild Cognitive Impairment. In *Lingue e Linguaggio*, 1/2021, 3-31, doi: 10.1418/101111

Gagliardi, G., Tamburini, F. (2022). The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5234–5242, Marseille, France. European Language Resources Association.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B. & Gorno-Tempini, M.L. (2014) Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. In *Cortex*, 55, 122–129.

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

Hernández-Domínguez, L., Ratte, S., Sierra-Martinez, G. et al. (2018). Computer based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. In *Alzheimer's Dement*, 2018; 10: 260–8. doi: 10.1016/j.dadm.2018.02.004

Karr, J.E., Graham, R.B., Hofer, S.M. & Muniz-Terrera, G. (2018). When does cognitive decline begin? A systematic review of change point studies on accelerated decline in cognitive and neurological outcomes preceding mild cognitive impairment, dementia, and death. In *Psychol. Aging*, 33(2), 195–218. doi: 10.1037/pag0000236

Kavé G., Dassa A. (2017). Severity of Alzheimer's disease and language features in picture descriptions In *Aphasiology*, doi: 10.1080/02687038.2017.1303441

Kim, B.S., Kim, Y.B. & Kim, H. (2019). Discourse measures to differentiate between mild cognitive impairment and healthy aging. In *Front. Aging Neurosci*. 11(221). doi: 10.3389/fnagi.2019.00221

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O. & Derreumaux, A. (2015) Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dement Diagnosis. In *Assess Dis Monit*. 2015;1:112–124.

Krein, L., Jeon, Y., Miller Amberber, A. & Fethney, J. (2019). The assessment of language and communication in dementia: A synthesis of evidence. In *The American Journal of Geriatric Psychiatry*, 27(4), 363–377. doi: 10.1016/j.jagp.2018.11.009

Lanzoni, A., Fabbo, A., Basso, D., Pedrazzini, P., Bortolomiol, E., Jones, M. & Cauli, O. (2018). Interventions aimed to increase independence and well-being in patients with Alzheimer's disease. Review of some interventions in the Italian context. In *Neurology, Psychiatry and Brain Research*. Volume 30, 2018, 137-143, ISSN 0941-9500. doi: 10.1016/j.npbr.2018.10.002

Luz, S., de la Fuente, S.D. & Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. In *Proceedings of the LREC 2018 Workshop "Resources*

and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)", Kokkinakis D, ed. ELRA, Paris.

Luz, s., Haider, F., de la Fuente, S., Fromm, D. & Macwhinney, B. (2020). Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. *Proceedings of Interspeech*, 2020, 2172-2176, doi: 10.21437/Interspeech.2020-2571

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2021a). Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proceedings of Interspeech 2021*, Brno, Czechia, 30 August – 3 September 2021, 3780–3784. doi:10.21437/Interspeech.2021-1220

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm D. & MacWhinney, B. (2021b). Alzheimer's dementia recognition through spontaneous speech. In *Frontiers in computer science*, 3(2021): 780169.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M. & Christensen, H. (2019). Dementia Detection Using Automatic Analysis of Conversations. In *Computer Speech Lang*. 53, 65–79. doi:10.1016/j.csl.2018.07.006

Mitchell, A.J. (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. In *Journal of psychiatric research,* 43, no. 4(2009): 411-431. doi: 10.1016/j.jpsychires.2008.04.014

Monroe T, Carter M. (2012). Using the Folstein Mini Mental State Exam (MMSE) to explore methodological issues in cognitive aging research. Eur J Ageing. 2012 Jun 15;9(3):265-274. doi: 10.1007/s10433-012-0234-8

Nevler, N., Ash, S., Irwin, D.J., Liberman, M. & Grossman, M. (2019) Validated automatic speech biomarkers in primary progressive aphasia. In *Annals of Clinical and Translational Neurology*, 6(1), 4–14.

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D. et al. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (1659-1666).

Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P. & Soyiri, I.N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. BMC Bioinformatics, 18(1), 34. https://doi.org/10.1186/s12859-016-1456-0

Ostrand, R. & Gunstad, J. (2021). Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults. In *Journal of Geriatric Psychiatry and Neurology*. 2021;34(5):357-369. doi:10.1177/0891988720933358

Panza, F., Solfrizzi, V., Barulli, M.R., Santamato, A., Seripa, D., Pilotto, A. & Logroscino, G. (2015) Cognitive Frailty: A Systematic Review of Epidemiological and Neurobiological Evidence of an Age-Related Clinical Condition. In *Rejuvenation Res*. 2015 Oct;18(5):389-412. doi: 10.1089/rej.2014.1637

Petti, U., Baker, S. & Korhonen, A. (2020), A systematic literature review of automatic Alzheimer's disease detection from speech and language. In *Journal of the American Medical Informatics Association*, 27(11), 1784–1797. doi: 10.1093/jamia/ocaa174

Pope, C., Davis, B.H. (2011). Finding a balance: The Carolinas Conversation Collection. In *Corpus Linguistics and Linguistic Theory*, 7(1):143–161. doi: 10.1515/cllt.2011.007

Prins, R.S., Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. In *Aphasiology,* 18, 1075–1091. doi: 10.1080/02687030444000534

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020.

Roark, B., Mitchell, M. & Hollingshead, K. (2007). Syntactic complexity measures for detecting Mild Cognitive Impairment. In: Cohen, K.B., Demner-Fushman, D., Frieman, C., Hirschman, L. & Pestian, J. (a cura di), *Proceedings of the Workshop BioNLP 2007: Biological, translational, and clinical language processing. ACL – Association for Computational Linguistics*, 1–8.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. In *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2081–2089. doi.org/10.1109/TASL.2011.2112351

Shah, Z., Sawalha, J., Tasnim, M., Qi, S., Stroulia, E. & Greiner, R. (2021). Learning language and acoustic models for identifying Alzheimer's dementia from speech. In *Frontiers in Computer Science*, 3(2021): 624659. doi: 10.3389/fcomp.2021.624659

Soria Lopez, J.A., González, H.M. & Léger, G.C. (2019). Alzheimer's Disease. Handb Clin. Neurol. 2019, 167, 231–255. doi: 10.1016/B978-0-12-804766-8.00013-3

Szmrecśanyi, B.M. (2004). On operationalizing syntactic complexity. In Purnelle, G., Fairon, C. & Dister, A. (Eds.), *Proc. of the 7th International Conference on Textual Data Statistical Analysis*, Presses Universitaires de Louvain, 1031–1038.

Teng, E.L., Chui, H.C. (1987). The Modified Mini-Mental State (3ms) Examination. In *J Clin Psychiatry*. 1987;48(8):314-318.

Themistocleous, C., Eckerström, M. & Kokkinakis, D. (2018) Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural In *Networks. Front. Neurol*. 2018, 9, 1–10.

Themistocleous, C., Ficek, B., Webster, K., den Ouden, D-B., Hillis, A.E. & Tsapkini, K. (2021) Automatic Subtyping of Individuals with Primary Progressive Aphasia. In *Journal of Alzheimer's Disease*, 79(2021) 1185–1194.

Toledo, C.M., Aluísio, S.M., Santos, L.B., Brucki, S.M.D., Trés, E.S., Oliveira, M.O. et al. (2017). Analysis of Macrolinguistic Aspects of Narratives from Individuals with Alzheimer's Disease, Mild Cognitive Impairment, and No Cognitive Impairment. In *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 31–40. doi:10.1016/j.dadm.2017.08.005

Tombaugh, T.N., McIntyre, N.J. (1992). The mini-mental state examination: a comprehensive review [see comment]. In *J Am Geriatr Soc*., 40(9): 922–935.

Tsoi, K.K, Chan, J.Y., Hirai, H.W., Wong, S.Y. & Kwok, T.C. (2015). Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. In *JAMA Intern Med*. 2015 Sep;175(9):1450-8. doi: 10.1001/jamainternmed.2015.2152

Ulatowska, H.K., Allard, L., Donnell, A., Bristow, J., Haynes, S.M., Flower, A. et al. (1988). Discourse performance in subjects with dementia of the Alzheimer type. In H.A. Whitaker (a cura di), *Neuropsychological studies of nonfocal brain damage: Dementia and trauma*. New York: Springer. doi: 10.1007/978-1-4613-8751-0_4

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

Vigorelli P. (ed.) (2004). *La conversazione possibile con il malato Alzheimer*. Milano: FrancoAngeli.

Vigorelli P. (2010). *The ABC Group for caregivers of persons living with dementia: self-help based on the Conversational and Enabling Approach. Non-pharmacological therapies in dementia*; 3: 271-286.

Vigorelli P. (2011). *L'approccio capacitante. Come prendersi cura degli anziani fragili e delle persone malate di Alzheimer*. Milano: FrancoAngeli.

Vigorelli P. (2021). *Dialoghi imperfetti. Per una comunicazione felice nella vita quotidiana e nel mondo Alzheimer*. Milano: FrancoAngeli.

Weiner, J., Schultz, T. (2016). Detection of intra-personal development of cognitive impairment from conversational speech. In *Speech Communication*, 12. ITG Symposium, Paderborn, Germany, 1-5.

Weiner, J. Frankenberg, C., Telaar, D., Wendelstein, B., Schröder, J. & Schultz, T. (2016). Towards Automatic Transcription of ILSE – an Interdisciplinary Longitudinal Study of Adult Development and Aging. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), 2016.

Yeung, A., Iaboni, A., Rochon, E. et al. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia In. *Alz Res Therapy,* 13, 109(2021). doi: 10.1186/s13195-021-00848-x