Computer Speech & Language 65 (2021) 101113

Contents lists available at ScienceDirect

Computer Speech & Language

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

Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia

Laura Calzà[a], Gloria Gagliardi[b], Rema Rossini Favretti[a], Fabio Tamburini[*,a]

[a] Alma Mater Studiorum - Università di Bologna, Italy
[b] Università degli Studi di Napoli "L'Orientale", Italy

A R T I C L E I N F O

Article History:
Received 10 December 2019
Revised 27 February 2020
Accepted 15 April 2020
Available online 29 June 2020

Keywords:
Mild Cognitive Impairment
Dementia
Language and speech analyses
NLP Techniques
Linguistic bio-marker

A B S T R A C T

Almost 50 million people are living with dementia in 2018 worldwide, and the number will double every 20 years. The effectiveness of existing pharmacologic treatments for the disease is limited to symptoms control, and none of them are able to prevent, reverse or turn off the neurodegenerative process that leads to dementia; therefore, a prompt detection of the "disease signature" is a key problem, in order to develop and test new drugs and to support the management of clinical and domestic context. Recent studies showed that linguistic alterations may be one of the earliest signs of the pathology, years before other neurocognitive deficits become evident. Traditional tests fail to identify these slight but noticeable changes; whereas, the analysis of spoken language productions by Natural Language Processing (NLP) techniques can ecologically and inexpensively identify minor language modifications in potential patients.

This interdisciplinary study aims at quantifying and describing alterations of linguistic features due to cognitive decline and build an automatic system for early diagnosis and screening purpose. To this aim, we enrolled 96 participants: 48 healthy controls and 48 impaired subjects. Of the latter, 32 was diagnosed with Mild Cognitive Impairment and 16 with early Dementia (eD). Each subject underwent a brief neuropsychological screening, and samples of semi-spontaneous speech productions was collected by means of three elicitation tasks. Recorded sessions were orthographically transcribed, PoS tagged and parsed building two different corpora: in the first we kept the automatic annotations, while in the second the transcripts were manually corrected in order to remove all mistakes. A multidimensional parameter computation was performed on the data, taking into consideration a set of 87 acoustical, rhythmical, morpho-syntactic and lexical feature as well as some readability indexes and demographic information. After these preparatory steps, some automatic classifiers were trained to distinguish healthy controls from MCI subjects employing two different algorithms, Support Vector (SVC) and Random Forest Classifiers (RFC). Our system was able to distinguish between controls and MCI subjects exhibiting high F1 scores, around 75%, thus it seems to be a promising approach for the identification of preclinical stages of dementia.

*Corresponding author at FICLIT, via Zamboni, 32, I-40126, Bologna, Italy.
E-mail address: fabio.tamburini@unibo.it (F. Tamburini).

https://doi.org/10.1016/j.csl.2020.101113
0885-2308/

1. Introduction

1.1. Cognitive decline as a growing public health concern

The number of people who are diagnosed with dementia is growing rapidly in western countries: almost 50 million people are living with dementia in 2018 worldwide, and the number will grow to 152 million by 2050. Rising life expectancy is contributing to rapidly boosts this number: through meta-analysis of the available evidence, Alzheimer's Disease International Association estimates over 9.9 million new cases of dementia each year, one new case every 3 seconds on average (Prince et al., 2015; Patterson, 2018).

The management of this increased number of individuals at risk of suffering from Cognitive Impairment is a big challenge for health-care systems: whilst the global societal cost of the pathology is barely feasible (US$ 818 billion, 1.09% of the global Gross Domestic Product), existing medications for the forms of clinically evident dementia, in particular for the Alzheimer's disease (AD), show minimal efficacy both on the cognitive and the functional manifestations of this ravaging condition.

However, the neurodegenerative process leading to dementia is thought to begin much earlier than the clinical symptoms: this long "preclinical" or "prodromal" phase, a grey area between normal aging and pathological cognitive functioning, would provide a key opportunity for pharmacological treatment development and therapeutic intervention (Calzà et al., 2015; Epelbaum et al., 2017; Ritchie et al., 2017). Customized interventions at early stages of the disease might reduce not only the economic impact of health spending, but also the emotional burden for patients and their caregivers. An adequate and timely risk identification may also allow for the implementation of preventive measures such as dietary, lifestyle and neuroprotection precautions, playing an important role in delaying the onset of the pathology.

Nevertheless, the problem of diagnosis in cognitive decline and frailty still remains a hot topic: there is an extensive literature and a considerable body of evidence on the possibility of early diagnosis of Alzheimer's and other types of dementia, but presymptomatic diagnosis raises both theoretical issues and ethical concerns (Calzà et al., 2015).

Individuals with dementia manifest alterations in various cognitive domains: memory, attention, executive functioning, visuo-spatial skills, perceptual speed and, last but not least, language. Many assessment tools have been proposed over recent years, but the commonest screening instruments (e.g. the "Mini Mental State Examination" (Folstein et al., 1975)) are largely inadequate for detecting early changes in cognition. It would be crucial to have high sensitive and specific psychometric tests, suitable for a low-cost and large-scale use. Several initiatives and studies are in progress (Mortamais et al., 2017), but, at the moment, the role of these traditional instruments is puzzling: although idoneous to detect evident dementia cases, they are much less effective in order to track down the prodromal phase of cognitive frailty, such as the condition of Mild Cognitive Impairment (MCI) (Petersen, 2011).

1.2. Quantitative linguistic methods and NLP techniques for cognitive frailty screening

Among the cognitive areas that may reveal early signs of decline, language has been subject of growing interest, becoming an established topic of research. Considerable evidence is available for suggesting that linguistic deficits are present in several neurodegenerative diseases (e.g. see Boschi et al., 2017 for a review); that is especially the case with dementia, where language disruption is a common finding both at the earliest stages and in full-blown pathology.

Although episodic memory impairment is the main symptom of AD, a progressive language disorder is usually found as well; but unlike aphasias, that are due to focal brain damage, verbal deficits usually occur in the context of multiple cognitive impairments (Forbes-McKay et al., 2013). Patients show, among other signs, a decline of lexical semantic knowledge, with word-finding problems (i.e. anomia and semantic paraphasias), sentence comprehension deficits, verbal fluency decrease and low content density (Catricalà et al., 2015; de Lira et al., 2011; Drummond et al., 2015; Fraser et al., 2016; Jarrold et al., 2014). At the phonetic level, low speech rate and an high number of hesitations have also been reported (Hoffmann et al., 2010; Sajjadi et al., 2012). Morpho-syntactic processing tends to be relatively preserved in the early course of the disease (Altmann et al., 2001; Cuetos et al., 2007; Sajjadi et al., 2012): nevertheless, a number of studies have showed that sentence structure is correct but reduced (Ferris and Farlow, 2013; Fraser et al., 2016; Kemper et al., 1993; Yancheva et al., 2015), and a greater number of inflectional errors in AD patients than in healthy persons has also been observed (Altmann et al., 2001; Cuetos et al., 2007). Deficits affect the pragmatic level too, namely referential/temporal cohesion, coherence and discourse planning (Carlomagno et al., 2005; Chapman et al., 2002; Drummond et al., 2015; March et al., 2006; Ripich et al., 2000). With the progression of AD pathology, linguistic symptoms become pervasive, showing a full breakdown of speech comprehension and verbal production restricted to echolalia and stereotypy (Ferris and Farlow, 2013).

To summarise, the progress of the pathology parallels with a progressive simplification of the language productions; this can be considered the first guiding line we followed searching for linguistic features able to distinguish among the various degrees of pathology for Italian speakers.

A progressive loss of specific language functions with relative sparing of other cognitive domains (such as memory of daily events, visual-spatial skills and behavior) marks out Primary Progressive Aphasia (PPA) (Mesulam, 2001; 2003): as a matter of fact, PPA is diagnosed when all major limitations in activities of daily living can be attributed to a language impairment for at least two years after the onset. Three subtypes are currently recognized: non-fluent/agrammatic variant PPA, semantic variant PPA, and logopenic variant PPA, each of which exhibits peculiar patterns of brain atrophy and linguistic features (Gorno-Tempini et al., 2011). People with non-fluent PPA show major impairments at the phonetical, phonological and syntactic levels: they usually present agrammatism in language production (i.e. omission of grammatical morphemes), effortful and halting speech with articulatory errors ("apraxia of speech"), disruption of prosody and impaired comprehension of complex sentences (i.e. negative passive and object relative clause) against spared single-word decoding and object knowledge. Severe anomia and single-word comprehension deficits, especially for low-frequency items (es. "zebra" vs the more familiar "horse"), are the core features of semantic variant PPA (also known as "Semantic Dementia"). These symptoms represent the earliest markers of a widespread conceptual knowledge degradation. People suffering from the logopenic PPA present impaired single-word retrieval and sentence repetition deficits. Word-finding problems bring about slow speech rate, but lack of frank agrammatism and preservation of speech articulation help in distinguishing it from other subtypes.

Language disruption is not a core feature of Dementia with Lewy Bodies (LBD), but quite the opposite (Ash et al., 2011; Delbeuck et al., 2013; Grossman et al., 2012). However, naming and verbal fluency impairment, due to disturbed executive functioning, have been extensively reported. In addition, alterations have been described both at the phonetic (e.g. speech rate, articulation errors) and pragmatic level (e.g. narrative organization, coherence and topic maintenance).

Although there is a lot of empirical evidence about language disruption in AD, PPA and LBD, less knowledge has been accumulated about language disorder in preclinical stages. Reviewing the literature on the topic, verbal impairments in MCI seem to parallel those found in early/moderate stage Dementia (Taler and Phillips, 2008): deficits are reported for verbal fluency, naming and semantic knowledge, even if pragmatic skills seem to be the most affected; it is also well documented that these discourse alterations (i.e. semantically impoverished discourse that lacks in coherence) may be one of the earliest signs of the pathology, often measurable years before other cognitive deficits become apparent. Some longitudinal retrospective studies have already demonstrated that linguistic features could act as a prodromic marker of cognitive dysfunctions: for example, the Nun study (Snowdon, 2003), the Iris Murdoch study (Garrard et al., 2005) or the Harold Wilson project (Garrard, 2009). The investigation of this domain seems to be promising, both for early diagnosis and dementia large-scale screenings. The traditional evaluation of the linguistic functions is performed by means of pencil-and-paper or computer-assisted tests: it is usually made up of verbal fluency (semantic and phonemic), visual confrontation naming, comprehension, repetition (of words and/or sentences), writing and communicative skill assessment. However, not only conventional language tests show low sensitivity, but they do not allow to explore many other aspects of language, both at the segmental and suprasegmental level (e.g. prosody and rhythm): albeit sporadic significant differences between the MCI and normal elderly participants have been identified with standardized instrument, their clinical use is still unreliable (Filiou et al., 2019; Szatloczki et al., 2015; Taler and Phillips, 2008).

During the last few years, new sophisticated techniques from Natural Language Processing (NLP) have been used to analyse written texts, clinically elicited utterances and spontaneous speech, in order to identify signs of psychiatric or neurological disorders and to extract automatically derived linguistic features for pathologies recognition, classification and description. Computational methods have been already successfully applied to the study of linguistic cues of cerebral functional disorders, both in the case of language modifications and disruption associated with depression (Jiang et al., 2017; Stasak et al., 2019), focal brain lesions (Fergadiotis and Wright, 2011), Parkinson's disease (Arias-Vergara et al., 2018; Benba et al., 2016; Sztahó and Vicsi, 2016; Upadhya et al., 2019) and for detecting dementia prodroms (MCI) (dos Santos et al., 2017; Matsuda Toledo et al., 2018; Meilán et al., 2018; Roark et al., 2007, 2011; Satt et al., 2013; Tóth et al., 2018; Vincze et al., 2016; Wang et al., 2019) or the different associated pathologies, like Alzheimer's Disease (Chinaei et al., 2017; Fraser et al., 2016; Jarrold et al., 2014; López-de-Ipiña et al., 2015; Yancheva and Rudzicz, 2016; Sirts, Piguet, Johnson), PPA (Fraser et al., 2014) and Fronto-Temporal Dementia (Jarrold et al., 2014).

While neuropsychological tests and structured evaluations have a relevant impact on the naturalness of the subject's responses, the analysis of natural spoken language productions could allow to ecologically and almost inexpensively pinpoint language modifications in potential patients even by primary care physicians.

Considering the cited literature globally, there are two main aspects that these works analyse trying to face the managed problem:

- the introduction of a proper set of linguistic features able to differentiate subjects that could present a pathological situation from controls. In this wide panorama, some studies extract the features manually from speech or written productions, while others try to devise NLP systems able to extract such features automatically. Almost all the cited works apply proper statistical significance test in order to identify the most promising linguistic indicators in correlation with the pathology studies.
- having defined a set of relevant features, some of the cited studies try to build automatic systems able to identify the pathology. They typically use common Machine Learning techniques, such as Support Vector Machines, Neural Networks, K-Nearest Neighbor, etc., to build such automatic tools obtaining different classification performances, in terms of accuracy or precision/recall/F-score.

1.3. Aim of the study

The paper presents a novel system for the identification of cognitive frailty at a very early stage by processing spontaneous Italian language productions. The final goal of this project regards the development of an instrument to be used at General Practitioner level, for frequent, low-cost and non-intrusive cognitive decline screening and cognitive status monitoring. In order to devise such a computational system it is necessary to study and compute a large set of linguistic features potentially able to distinguish between healthy controls and MCI subjects in a reliable way for the Italian language and Italian speakers. For this reason, a large set of indexes identified as relevant in studies devote to other languages, as well as some linguistic parameters never used for this task or specifically computed for Italian, have been identified and considered into this work, and proper procedures for measuring them starting from spoken productions have been developed.

Over recent years, there has been a huge increase in the number of scientific papers on the topic but, at the time of writing, we are not aware of any study specifically devoted to Italian performing a similar kind of automatic analysis. This is not a marginal issue, since typological peculiarities (e.g. morphological structure) may strongly affect the reliability of the features on different languages, limiting the comparability and transferability of results. This pilot study represents the first step in the direction of creating automatic tools able to support practitioners taking the very burdensome, but fundamental, decision if a subject presents real symptoms of cognitive decline, deserving further diagnostic investigation. With respect to previous works (e.g. Beltrami et al., 2018), this paper represent our very first attempt to move from the theorical linguistic profiling of MCI population to its actual automatic identification.

2. Material and methods

2.1. Corpus design and interview recording

We enrolled 96 participants, 48 healthy controls ("control group", CG) and 48 subjects with cognitive decline ("pathological group", PG). All of them provided informed and written consensus. The sample will be balanced by sex, age (range 50−75) and education (primary school with great intellectual stimulation throughout the life span or junior high school; high school; academic degree). The PG included 48 participants from two outpatient clinical services involved in-care and diagnostic evaluation of cognitive disorders and dementia.

It refers to two categories:

1. Mild Cognitive Impairment (MCI): it causes cognitive changes that are serious enough to be assessed with neuropsychological assessment, but not so severe to interfere with everyday activities. In order to provide a good balance among clinical phenotypes of the disease, the sample is further divided into:
   - amnestic MCI single domain (a-MCI; 16 subjects): patients who show an isolated memory deficit;
   - multiple domain MCI (md-MCI; 16 subjects): in these individuals two or more cognitive abilities are affected (memory can be engaged or not).
2. Early Dementia (e-D; 16 subjects): these patients are affected by cognitive deficits which partially influence everyday life.

Table 1 summarises the main characteristics of our cohort. Complete information about the subjects sampling and the corpus building can be found in Beltrami et al. (2018).

All the participants of the CG and PG were requested to complete the anamnestic interview (including anagraphic data, information about occupation/retirement, children, familiarity with ne``urodegenerative pathologies, clinical history and pharmacotherapy). While CG only went through the conventional cognitive battery, all the PG participants underwent a complete neuropsychological evaluation (comprehensive of a clinical interview, an assessment of cognitive, emotional and behavioral features, a self care abilities analysis and an interview with a family member, whenever possible), a neurological assessment and other medical examinations planned in the diagnostic work-up. The cognitive battery was composed of those tests which are most used in the clinical practice to assess cognitive decline (Velayudhan et al., 2014; Tsoi et al., 2015), with an Italian standardization and short administration time.

Table 1
Description of the cohort: inclusion criteria (i.e. MMSE Raw score (Folstein et al., 1975; Measso et al., 1993) and MoCA (Conti et al., 2015; Nasreddine et al., 2005)), age and education. No age differences were observed between the subgroups (non-parametric Kruskal-Wallis tests with Dunn's multiple comparison, $p > 0.05$); on the contrary, the level of education of the eD group is significantly lower than Healty Controls (p-value = 0.0171).

	Healty Controls	MCI subjects	eD subjects
inclusion criteria	MMSE $\geq$ 24; MoCA $\geq$ 18 No neurological pathologies, sensory impairment or intellectual disability No familiarity with dementia	MMSE $\geq$ 18 No problem in activities of daily living	MMSE $\geq$ 18 Need of support for one or more activities of daily living
age	61.60 $\pm$ 6.93	64.34 $\pm$ 7.33	66.38 $\pm$ 6.70
education	13.00 $\pm$ 3.92	11.28 $\pm$ 4.35	9.38 $\pm$ 4.01 *

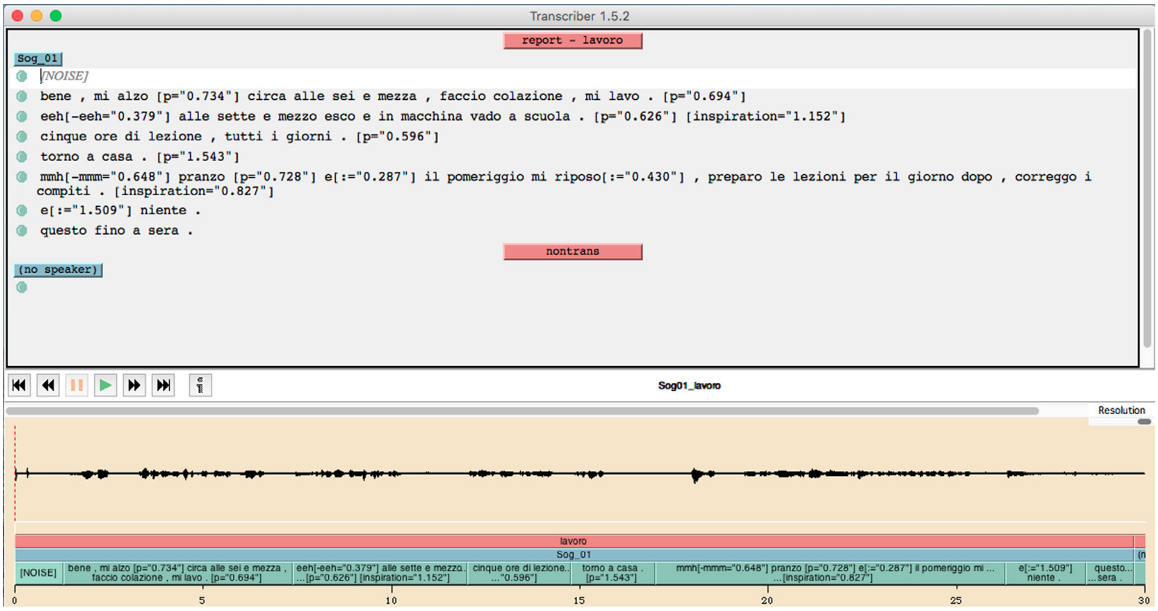

Fig. 1. Orthographic transcription of a short speech sample with the Transcriber software. From the top: text editor for speech turn annotation and segmentation; button bar for signal playback; waveform; segmentation lines (synchronized with the signal).

After the traditional neuropsychological assessment, we recorded the semi-spontaneous speech of the subjects during the execution of three mnemonic/linguistic tasks, elicited by these input sentences:

- ”Could you please describe this picture?” (the picture illustrated a living room with some characters carrying out certain actions, from Ciurli et al., 1996); Task “*FIGURE*”;
- ”Could you please describe your typical working day?”; Task “*WORK*”;
- ”Could you please describe the last dream you remember?”; Task “*DREAM*”.

Speech samples have been recorded in a quiet room with an Olympus-Linear PCM Recorder LS-5 (in WAV files; 44.1KHz, 16 bit) placed on a table in front of the subject.

2.2. Corpus transcription and annotation

From the digital recordings of the subjects’ interviews we realized two distinct corpora. The first is a Manually Checked Corpus (MCC) built by either fully manual or semi-automatic techniques: in both cases the data in the corpus, from the speech transcriptions to the whole set of linguistic annotations, are completely reliable and manually checked at any level. The second corpus is an Automatically Annotated Corpus (AAC), meaning that all the steps, speech transcriptions and linguistic annotations, have been obtained by an automatic procedure applying different speech and NLP tools and not corrected in any way. The reasons for building such different resources are twofold: first we used MCC in order to investigate the relationships between linguistic features and the different subjects groups, in order to clarify if it is possible to define a set of features enabling a sufficiently reliable distinction between them by using automatic classifiers; for this preliminary stage we preferred to base our conclusions on reliable data, manually checked. Then we used the findings of the first stage and applied the same procedures on data annotated by using automatic tools, in order to get an idea about the decrease in performance exhibited by a fully automatic system.

Unfortunately, only 92 sessions over 96 have been processed due to recording quality problems, mainly excessive noises. These speech samples, collected from 4 control subjects, have been excluded from the analysis.

In current literature we can find a high number of studies demonstrating that it is easy to reliably identify subjects presenting a recognised stage of dementia from healthy controls (e.g. (Fraser et al., 2016; Jarrold et al., 2014; López-de-Ipiña et al., 2013, 2015; Sirts, Piguet, Johnson), but this is not useful, because once the clinical symptoms have been identified with certainty it is often too late to intervene in a proper way to contain the illness.

In this study we are mainly interested in developing proper procedures for the automatic identification of the early stages of the pathology, thus we will concentrate our analyses on the discrimination between control and MCI groups; therefore, our ultimate sample will be composed of 44 controls and 32 MCI subjects. Even if the results presented in this paper are mainly devoted to control/MCI distinction, all the interviews from the 92 subjects have been completely processed and annotated.

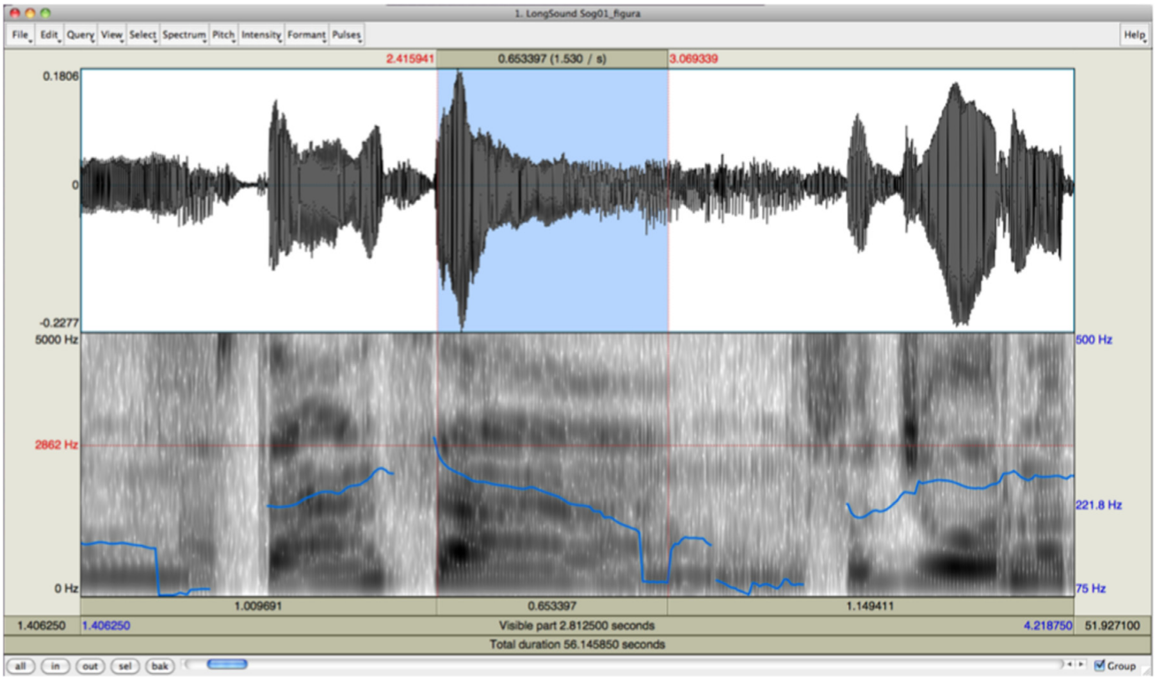

Fig. 2. Measuring the length of linguistic and extra-linguistic events with Praat.

2.2.1. Manually checked corpus - MCC

The speech samples have been manually transcribed by using *Transcriber*[1], a free tool for helping scholars to transcribe speech dialogues keeping track of different turns or the various linguistic phenomena in the audio samples (see Fig. 1 for an example of transcriptions with the cited tool). Output files are exported in an XML format with temporal alignment of the text to the signal and Unicode UTF-8 character encoding. The operating procedure is compliant with the annotation guidelines of the project, available to the transcribers.

The reference unit for the analysis of the speech flow is the utterance, defined by using pragmatic and prosodic (mainly intonational) criteria as "the linguistic counterpart to the speech act", the minimal linguistic entity that is pragmatically interpretable (Austin, 1962). The identification is performed through the perception and the detection of prosodic breaks (Cresti and Moneglia, 2018) acoustically correlating with F0 reset, final lengthening, drop in intensity, pause and initial rush in the subsequent prosodic unit (Cruttenden, 1986; Hirst and Di Cristo, 1998; Sorianello, 2006). One or more utterances performed without interruptions by a single speaker make up a "dialogic turn".

Orthographic transcription follows the conventions of written Standard Italian; in order to dispel any spelling doubts, the annotators referred to the "GRADIT" dictionary (De Mauro, 1999). During the transcription process a set of paralinguistic and extralinguistic phenomena (such as empty or filled pauses, disfluences, lapsus, hesitations/statterings, laughs, coughs, throat clearing sounds or noises) has been annotated as well. All labels were clearly marked in order to allow an easy removal of the annotations from the corpus and the reversion to the raw data (Leech, 2005). The duration of the verbal and non-linguistic events (in ms.) has been annotated too, gauging their temporal extension on the spectrogram by using the Praat speech processing tool[2] (Boersma, 2001) (see Fig. 2).

Table 2 shows means and standard deviations for the lengths of the texts uttered by controls and MCI subjects.

After the automatic tokenization of the transcription, the corpus has been enriched by adding linguistic information at the lexical and morphosyntactic levels: all the utterances have been automatically PoS-tagged, lemmatized and syntactically parsed with the dependency model used by the Turin University Linguistic Environment-TULE[3] (Lesmo, 2007), based on the TUT - Turin University TreeBank tagset (Bosco et al., 2000). Given that parser performance gets worse with transcripts, even more with pathological language, we decided to rely on carefully checked linguistic information, at least for MCC. To this end, all the annotations have been manually checked by one linguist, in order to remove the errors introduced by the automatic tagging. The revision has been made by using the *Dependency Grammar Annotator* - DGA opensource software[4] for an easy visualisation and correction of TULE mistakes at any level (see Fig. 3).

1 http://trans.sourceforge.net
2 http://www.fon.hum.uva.nl/praat/
3 https://github.com/alexmazzei/TULE
4 http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/

Table 2
Text length, in tokens, produced on the three tasks by healthy and MCI subjects, shown as mean (μ) and standard deviation (σ).

Task	Healty Subjects	MCI subjects
FIGURE	μ=184.75, σ=113.60	μ=121.91, σ=52.63
WORK	μ=286.32, σ=220.58	μ=238.66, σ=216.06
DREAM	μ=135.07, σ=100.72	μ=94.97, σ=66.96

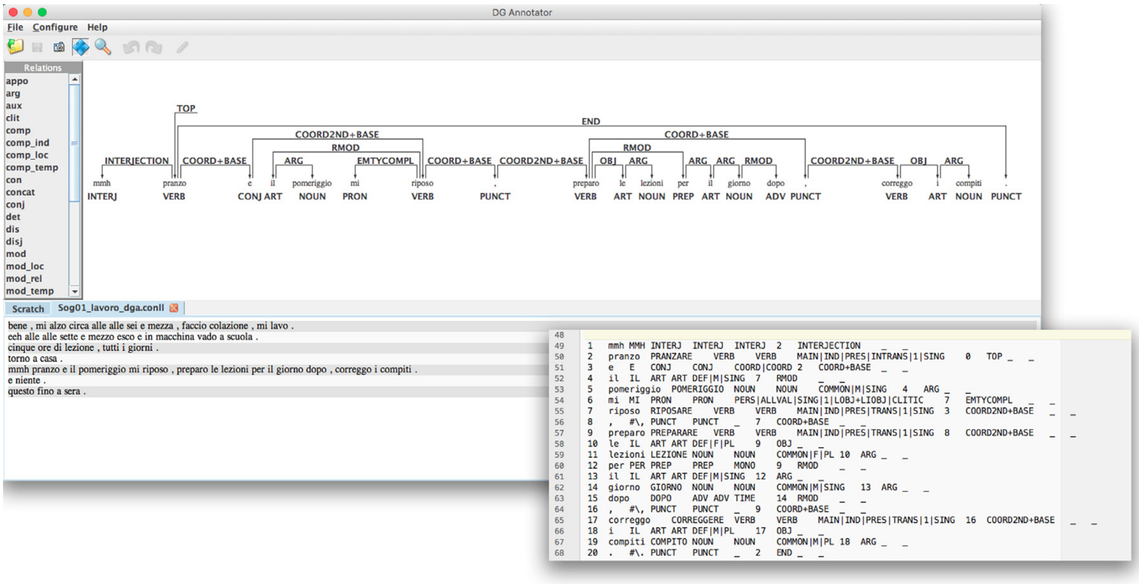

Fig. 3. Dependency graph as shown by DGA and full utterance annotation in CoNLL-U format (lemmas, Part-of-Speech and dependency relations).

2.2.2. Automatically annotated corpus - AAC

This second version of our dataset has been processed by using completely automatic procedures, both for obtaining the initial transcriptions of the interviews and for inserting the same linguistic annotation we described before for MCC.

We produced the speech transcriptions of the interviews by leveraging the Google Cloud Automatic Speech Recognition (ASR) system, well trained for Italian and able to produce reliable transcriptions for non pathological language.

In order to have an idea about the actual performance of this tool when applied to pathological language, we ran an evaluation experiment comparing the automatically-derived transcriptions with the manual counterparts: we obtained a Word Error rate (WER) of 27.78%. Considering the nature of our speech samples this WER can be seen as rather acceptable.

With regard to the insertion of linguistic annotations, mainly part-of-speech tagging, lemmatisation and dependency parsing, we relied on the same tool used for MCC, namely the TULE parser, we described in the previous section. Unfortunately, it is not possible running a simple evaluation to get a broad idea about the amount of errors also for these linguistic annotations, due to the inevitable misalignments between the transcriptions that do not allow PoS-tagging and parsing evaluations with standardised procedures.

2.3. Linguistic features

A multidimensional parameter analysis has been performed on the two corpora: the algorithms conduct a quantitative analysis of spoken texts, computing rhythmic, acoustic, lexical, morpho-syntactic and syntactic features. Both linguistic/stylometric indexes proposed in the literature and some new parameters have been tested, for a total of 87 variables. Age and Cognitive Reserve (CR), namely the ability to optimize and maximize performance through the recruitment of brain networks and/or compensation by alternative cognitive strategies (Nucci et al., 2012), are among the most important risk factors for Mild Cognitive Impairment (Mazzeo et al., 2019). So far, CR has been estimated by extremely heterogeneous methods and proxy measures: years of education, occupation, intelligence (IQ), leisure activity. In particular, the scientific community agrees on the role played by education in the cognitive decline both in normal aging and degenerative disease: even if its effect is rather difficult to isolate from other protective factors (e.g. socioeconomic status, quality of the social environment, awareness of health risks), scholarity have a huge influence on adult lifestyles. Nevertheless, scientific evidence for CR correction is lacking. Since information on age

Table 3
Acoustic Features.

Feature	Description	References
Silence segments duration	Silence segments of the signal identified using a voice activity detector (VAD). Mean (SPE_SILMEAN), median (SPE_SILMEDIAN) and Std. Deviation (SPE_SILSD) were taken into account.	Satt et al. (2013)
Speech segments duration	Speech segments of a signal identified using a voice activity detector (VAD). Mean (SPE_SPEMEAN), median (SPE_SPEMEDIAN) and Std. Deviation (SPE_SPESD) were taken into account.	Satt et al. (2013)
Temporal regularity of voiced segments	The measure captures the temporal structure of the voiced segments, providing information on the rate of change in the different spectrum bands. To calculate the temporal regularity of voiced segment durations, we used the sequence of the duration values, and calculated the real cepstrum of the sequence (SPE_TRVSD).	Satt et al. (2013)
Verbal Rate	The number of words in the sample divided by the Total Locution Time (i.e. speech time including pauses) (SPE_VR). $words/TLT$	Singh et al. (2001); Roark et al., 2011
Transformed Phonation Rate	The arcsine of the square root of the Phonation Rate. $\arcsin\sqrt{PR}$ Where PR is the phonation rate $PR = TPT/TLT$ TPT: total phonation time (i.e. speech time without pauses) TLT: total locution time (i.e. speech time including pauses). The arcsin transformation (or "angular transformation") provides a normally distributed measure within each participant group (SPE_TPR).	Singh et al. (2001); Roark et al., 2011
Standardized Phonation Time	The number of words in the sample divided by the total phonation time (i.e. speech time excluding pauses) (SPE_SPT). $words/TPT$	Singh et al. (2001); Roark et al., 2011
Standardized Pause Rate	The number of words in the sample divided by pauses (SPE_SPR). $words/pauses$	Singh et al. (2001); Roark et al., 2011
Root Mean Square energy	Physically, energy is a measure of "how much signal" exists at any one time, and it is used in continuous speech to detect voiced sounds, which have higher intrinsic energy than unvoiced segments. The energy of a signal is typically calculated by windowing the signal at a particular time, squaring the samples and taking the average. The square root of this result is the engineering quantity known as the root-mean square (RMS) value. Mean (SPE_RMSEM) and Std. Deviation (SPE_RMSESD) of the measures were taken into account.	López-de-Ipiña et al., 2013
Pitch	Pitch is the main acoustic correlate of tone and intonation, and the perceptual correlate of frequency; as a matter of fact, it depends on the number of vibrations per second produced by the vocal cords. Mean (SPE_PITCHM) and Std. Deviation (SPE_PITCHSD) were taken into account.	López-de-Ipiña et al., 2013
Spectral Centroid	The measure captures the perceptual brightness of a sound. It is obtained by evaluating the "centre of gravity" of the spectrum using the Fourier transform's frequency and magnitude information. Mean (SPE_SPCENTRM) and Std. Deviation (SPE_SPCENTRSD) were taken into account.	López-de-Ipiña et al., 2013
Higuchi Fractal Dimension	The feature describes the "complexity" of the signal. The algorithm measures fractal dimension (i.e. self-similarity, namely identical/similar structures repeating over a pattern) of discrete time sequences directly from time series. Mean (SPE_HFractDM) and Std. Deviation (SPE_HFractDSD) were taken into account.	López-de-Ipiña et al., 2013

and education is available in our clinical dataset, they have been added to the input variables of the classifiers, as "demographic features".

Tables 3–8 outline the complete list and the description of the features considered in this study.

2.4. Feature extraction and data processing

With regard to the parameters derived from the speech acoustics, we used the "SSVAD v1.0" Voice Activity Detector proposed by (Yu and Mak, 2011)[5], especially developed for interview speech, to segment the recordings and identify speech vs non-speech regions. Those segmentations were fundamental for computing some acoustic features like silence and speech segments duration. We relied also on a forced alignment system we developed by using the Kaldi-DNN-ASR package[6] trained on the APASCI

[5] http://bioinfo.eie.polyu.edu.hk/ssvad/ssvad.htm
[6] https://kaldi-asr.org/

Table 4
Demographic features.

Feature	Description	References
Age	Subject's age (NPT_AGE).	-
Scholarity	Subject's number of school years (NPT_SCHOOL).	-

Table 5
Readability Features.

Feature	Description	References
Text readability	It is a set of four readability features as computed by the READ-IT readability assessment tool: it computes a lexical-based index of readability (REA_BASE), a morpho-syntactic readability index (REA_MOSYN), a syntactic readability index (REA_SYNTAX) and a combination of the previous ones (REA_ALL).	Dell'Orletta et al. (2011)

Table 6
Rhytmic Features.

Feature	Description	References
Percentage of vocalic intervals	The proportion of vocalic intervals within the utterance, that is, the sum of vocalic intervals divided by the total duration of the utterance (RHY_%V).	Ramus et al. (1999)
Std. deviation of vocalic and consonantal interval durations	The standard deviation of the duration of vocalic and consonantal intervals within each utterance, noted as ΔV (RHY_DeltaV) and ΔC (RHY_DeltaC).	Ramus et al. (1999)
Pairwise Variability Index, raw and normalized	This rhythm metric takes into account the temporal succession of the vocalic and consonantal intervals instead of joining all the values and calculating the standard deviation. It is based on a pairwise comparison of the durations of either two vocalic (RHY_VnPVI) or consonantal (RHY_CrPVI) intervals, therefore expressing the level of variability in consecutive measurements. Raw Pairwise Variability Index (rPVI): $rPVI = \left[\sum_{k=1}^{m-1} \lvert d_k - d_{k+1} \rvert / (m-1)\right]$ where m is number of intervals, vocalic or intervocalic, in the text and d is the duration of the interval. Normalised Pairwise Variability Index (nPVI): $nPVI = 100 \cdot \left[\sum_{k=1}^{m-1} \lvert \frac{d_k - d_{k+1}}{(d_k - d_{k+1})/2} \rvert / (m-1)\right]$	Grabe and Low (2002)
Variation coefficient for ΔV and ΔC.	A variation coefficient (*varco*) is a value describing relative variation. $Varco\Delta C$ (RHY_VarcoC) is calculated as the percentage of the ΔC of the average duration of intervals (*meanC*); analogously, $Varco\Delta V$ (RHY_VarcoV) is calculated as the percentage of the ΔV of the average duration of intervals (*meanV*). $Varco\Delta C = \Delta C \cdot 100 / meanC Varco\Delta V = \Delta V \cdot 100 / meanV$	Delwo (2006)

Italian Corpus[7]: the forced aligner enabled us to obtain the temporally aligned phonetic transcriptions needed to compute the rhythmic features listed in Table 6.

All input features have been z-score normalised: for any feature X we computed its mean (μ_X) and standard deviation (σ_X) across the dataset and applied the formula $Z_X = X - \mu_X / \sigma_X$. This is a standard procedure for compacting the data features around zero helping the ML classifiers to achieve better performance. In a previous work (Beltrami et al., 2018) we carefully analysed the correlations between all the linguistic features previously described, computed on the MCC, and the neuropsychological test taken to define the gold standard and to produce the final subjects classifications.

We performed also a statistical analysis by computing the significance for each feature by using the Kolmogorov-Smirnov non-parametric test. We chose such kind of hypothesis testing technique, compared with the *t*-test or the Wilcoxon-Mann-Whitney test, because of the small size of our corpus. Table 9 outlines the different levels of significance for the considered linguistic features. Acoustic features (SPE_*), directly derived from the recordings, play a central role in distinguishing the two classes of subjects (CG vs PG), being almost all highly significant. With regard to lexical (LEX_*) and syntactic (SYN_*) features, some of them are very significant and thus interesting, while the other families of indexes (rhythmic - RHY_*, readability - REA_* and demographic - NPT_*) seem not so relevant, or only slightly relevant, for supporting the classification process.

Before describing the classification experiments and the obtained results we have to briefly discuss an important point. Most of the works in the relevant literature tend to study a group of linguistic features from a statistical point of view, and then build the automatic classifier by only using the significant feature or reducing the number of considered features by applying feature selection/dimension reduction techniques (such as Minimum Redundancy Maximum Relevance). This is usually done on the same dataset used for training the classifier introducing, in our opinion, a bias on classifiers' evaluation.

In this work we did not apply any feature selection method and did not select the feature looking at their statistical significance (we made only some experiments reported in Section 3 for comparison purpose). We built our classifiers by using the whole set of listed features and let the automatic classifiers the task of identifying which complex combination of features achieve the best results. In this way we were certain not to introduce any kind of bias into the evaluation procedure, and thus to

[7] http://catalogue.elra.info/en-us/repository/browse/ELRA-S0039/

Table 7
Lexical Features.

Feature	Description	References
Content Density	The ratio of open-class words to closed-class words. The measure is calculated over Part of Speech tags, where open-class words are nouns, verbs, adjectives, adverbs; the rest are considered closed-class words (LEX_ContDens). $ContentDensity = OCW/CCW$	Roark et al., 2011
Part-of-Speech rate	This class of features investigates the average rate of occurrence for each part-of-speech (PoS) category: Adjectives, Adverbs, Articles, Conjunctions, Interjections, Nouns, Numerals, Prepositions, Pronouns, Verbs (LEX_PoS_*). $e.g.: Adjectives/words$	Holmes and Singh (1996); Bucks et al. (2000)
Reference Rate to Reality	The ratio of the total number of nouns to the total number of verbs (LEX_RefRReal). $RefRReal = Nouns/Verbs$	(Vigorelli, 2004)
Personal, Spatial and Temporal Deixis rate	The feature probes the rate of deictic expressions in the spoken text (i.e. linguistic elements that point to the time, place, or situation in which a speaker is speaking; in other words, their denotational meaning varies depending on extralinguistic context). The main types of deixis are: (a) Person deixis (e.g. I, you, we, me, mine, yours...) (LEX_PDEIXIS); (b) Place deixis (e.g. here, there, this, that...) (LEX_SDEIXIS); (c) Time deixis (e.g. now, today, tomorrow, soon...) (LEX_TDEIXIS). $e.g.: PersonDeixis/words$	March et al., 2006; Cantos-Gòmez (2009)
Relative pronouns and negative adverbs rate	The rate of Relative Pronouns (e.g. who, whose...) (LEX_RPRO) and Negative Adverbs (e.g. not, neither...) (LEX_NEGADV) in the spoken text.	
Lexical Richness: Type-Token Ratio, W - Brunet's Index and R - Honor's Statistic	This class of measures quantifies the richness of vocabulary/lexical diversity. *TTR, Type-Tokes Ratio*: the ratio of the number of different words (vocabulary - V) to the total text length. TTR is dependent on the text size: it is bigger when texts are small and decreases as the texts get larger (LEX_TTR). *W, Brunet's Index*: it quantifies lexical richness without being sensitive to text length. It is calculated according to the following equation: $W{=}_N V^{(-.165)}$ where N is the total text length and V is the total vocabulary used by the participant. This measure generally varies between 10 and 20. The lower the value, the richer the speech (LEX_BrunetW). *R, Honoré's Statistic*: calculates lexical richness by highlighting the proportion of words that are used only once with reference to the total number of words in the text: the larger the number of words used by a speaker that occur only once (hapax legomena), the richer the lexicon. $\frac{R{=}100logN}{(1-V1/V)}$ where V1 is the words spoken only once, V is the total vocabulary used and N is the total text length. High value of R suggests a rich vocabulary used by the speaker (LEX_HonoreR).	Holmes and Singh (1996); Brunet (1978); Honoré (1979)
Action Verbs rate	The metric probes the rate of action verbs (i.e. verbs referring to physical action, like to put, to run, to eat) in the spoken text (LEX_ACTVRB).	Gagliardi (2014)
Frequency-of-use tagging	Mean frequency-of-use weight among words extracted from the De Mauro's frequency list (LEX_DM_F).	De Mauro (2000)
Propositional Idea Density	Idea density is the number of expressed propositions (i.e. distinct facts or notions contained in a text) divided by the number of words. It is a measure of the extent to which the speaker is making assertions (or asking questions) rather than just referring to entities. In this feature, propositions correspond to verbs, adjectives, adverbs, prepositions, and conjunctions. Nouns are not considered to be propositions, as the main verb and all its arguments count as one proposition (LEX_IDEAD).	Snowdon et al., 1996; Roark et al., 2011
Mean Number of words in utterances	Mean number of words in the speech utterances (LEX_NW).	

be able to trust the system's performance as a good approximation of its generalisation abilities on new unseen data. In our opinion, given the relatively small size of the dataset, this is the safest way to design experiments not affected by any bias.

2.5. Automatic classifiers

As stated in previous sections, the long-term goal of this project is the construction of an automatic system able to help the practitioner in screening the linguistic status of all her/his patients in order to identify early signs of cognitive frailty and, in particular, MCI states. Designing proper subjects classifiers based on their speech productions in solving very simple and specific tasks could be a viable solution.

As a pilot study we followed this idea experimenting around the construction of such a classifier by using all the considered linguistic features as the input for some Machine Learning (ML) technique and training the system accordingly.

In the deep learning era the use of Deep Neural Networks (DNN) seems an obvious choice when building any kind of automatic classifier, but this is not the case: DNNs classifier training requires a large amount of annotated data that is not available, in general, when working with this kind of studies and certainly not available in our study. When data is scarce, other techniques proved to be equally useful allowing for the construction of reliable classifiers even if trained with small amounts of data. We

Table 8
Syntactic Features.

Feature	Description	References
Number of dependent elements linked to the noun	The feature explores Noun Phrase complexity, counting the number of dependent elements linked to the head (e.g. Adjectives, Relative clauses...). Mean (SYN_NPLENM) and Std. Deviation (SYN_NPLENSD) were taken into account.	
Global Dependency Distance	Given the memory overhead of long distance dependencies, the feature quantifies the difficulty in syntactic processing. Mean (SYN_GRAPHDISTM) and Std. Deviation (SYN_GRAPHDISTSD) were taken into account.	Roark et al., 2007, 2011
Syntactic complexity	Syntactic complexity is established by counting the linguistic tokens that can be considered to telltale signs of increased grammatical subordinateness and embeddedness, such as: subordinating conjunctions (e.g. because, since, as, when, that, etc.), WH-pronouns (e.g. who, whose, whom, which), verb forms, both finite and non-finite, noun phrases. Because subordinators and WH-pronouns are the most straightforward indicators of increased embeddedness (and thus of high complexity), these features are weighted more heavily than verbal forms and noun phrases (SYN_ISynCompl). $\frac{(2 \cdot conj + 2 \cdot pron + nouns + verbs)}{word}$	Szmrecsányi (2004)
Syntactic embeddedness: maximum depth of the structure	Syntactic complexity is also assessed by evaluating the “embeddedness”, i.e. the maximum ”dep” of the structure. Mean (SYN_MAXDEPTHM) and Std. Deviation (SYN_MAXDEPTHSD) were taken into account.	
Utterance length	Mean Length of utterance corresponds to the average number of words for utterance. It is calculated by counting the number of words in each utterance divided by the total number of utterances. Mean (SYN_SLENM) and Std. Deviation (SYN_SLENSD) were taken into account.	

experimented with two different supervised models implemented in the scikit-learn[8] python package: C-Support Vector Classifiers (SVC) (Cortes and Vapnik, 1995) and Random Forest Classifiers (RFC) (Tin Kam Ho, 1998). For a detailed description of these common ML techniques see also (Flach, 2012).

SVCs are machine learning techniques that, given a set of training examples each marked as belonging to one or the other of two categories, builds a model that assigns examples to one category or the other dividing the categories by a gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. In addition to performing linear classification, SVCs can efficiently perform a non-linear classification using what is called the “kernel trick”, implicitly mapping their inputs into high-dimensional feature spaces.

RFCs is an ensemble learning method for classification that operates by constructing a set of decision trees at training time and outputting the class that is the mode of the classes of the individual trees, correcting the decision trees’ habit of overfitting to their training set.

For our experiments we employed a Nested Leave One-Speaker-Out Cross Validation (NLOSOCV) (Krstajic et al., 2014) both for model selection, by optimising the hyperparameters over the validation set, and for model assessment evaluating the system on the test set. Given the limited size of the dataset we chose the “leave one-speaker out” solution to maximise the amount of data in the training set averaging the results obtained from the different folds.

The performance achieved by the system will be evaluated in terms of macro-averaged F1-score. Given the complexity of the whole evaluation procedure and the influence of the starting random point for each training/validation/test procedure on the evaluation results, we ran 10 different NLOSOCV procedures for any result computing mean and standard deviation of the F1-scores over the 10 runs.

With regard to the systems developed using C-Support Vector Classifiers we used RBF Kernels optimising the model parameters with a grid search, directly embedded into the NLOSOCV procedure, in the intervals $\gamma \in [0.0005, 1.0]$ and $C \in [0.0001, 100]$ where γ is the parameter of the RBF kernel and C is the SVC penalty parameter of the regularisation term. Random Forest Classifiers need only to determine a single parameter n_{trees} representing the number of trees in the forest; we optimised this parameter in the interval $n_{trees} \in [5, 100]$ during the NLOSOCV procedure.

3. Results and discussion

Table 10 outlines the results obtained by the different classifiers on the two corpora, MCC and AAC, expressed as macro-averaged F1-score, for all tasks together (first column) and for any single task.

The first observation we can make by looking at the results regards the good performance of the classifiers when trained on manually checked data (MCC) and all tasks. F1-scores well above 74% seems very encouraging and certainly far from a random classification (50%). The performance on single tasks drops considerably. In our opinion, there are two possible explanations for this fact:

[8] https://scikit-learn.org

Table 9
Linguistic features considered in this study (see Tables 3–8 for descriptions and abbreviations): means and standard deviations of their normalised values, distinguishing between controls and MCI subjects, and statistical significance at the Komolgorov-Smirnov (KS) test (*: 0.01<p-value<0.05, **: 0.001<p-value<0.01, ***: p-value<0.001).

Feature Type	Feature	MCI (PG)		Controls (CG)		KS	
		μ	σ	μ	σ	p-value	Signif.
Acoustic	SPE_SILMEAN	0.34	1.20	-0.25	0.73	<0.001	***
	SPE_SILMEDIAN	0.26	1.02	-0.19	0.95	<0.001	***
	SPE_SILSD	0.26	1.00	-0.19	0.96	<0.001	***
	SPE_SPEMEAN	-0.49	0.68	0.36	1.05	<0.001	***
	SPE_SPEMEDIAN	-0.41	0.70	0.30	1.08	<0.001	***
	SPE_SPESD	-0.44	0.55	0.32	1.12	<0.001	***
	SPE_TRVSD	0.01	0.96	-0.00	1.03	0.666	
	SPE_VR	-0.20	1.16	0.14	0.84	0.005	**
	SPE_TPR	-0.50	1.07	0.36	0.77	<0.001	***
	SPE_SPT	0.35	1.38	-0.25	0.45	<0.001	***
	SPE_SPR	-0.41	0.72	0.30	1.07	<0.001	***
	SPE_RMSEM	-0.24	0.96	0.18	1.00	0.015	*
	SPE_RMSESD	0.05	1.06	-0.04	0.96	0.428	
	SPE_PITCHM	0.16	1.00	-0.11	0.99	0.137	
	SPE_PITCHSD	0.13	1.06	-0.10	0.94	0.407	
	SPE_SPCENTRM	0.36	1.21	-0.26	0.71	<0.001	***
	SPE_SPCENTRSD	0.11	1.19	-0.08	0.83	0.232	
	SPE_HFractDM	0.40	0.95	-0.29	0.93	<0.001	***
	SPE_HFractDSD	-0.23	1.05	0.17	0.93	<0.001	***
Demographic	NPT_AGE	0.19	1.02	-0.14	0.96	0.015	*
	NPT_SCHOOL	-0.24	1.03	0.17	0.94	0.036	*
Readability	REA_BASE	-0.29	1.25	0.21	0.71	0.017	*
	REA_MOSYN	-0.17	1.03	0.12	0.96	0.013	*
	REA_SYNTAX	-0.20	1.10	0.15	0.90	0.074	
	REA_ALL	-0.23	1.14	0.17	0.85	0.066	
Rhytmic	RHY_%V	-0.03	1.11	0.02	0.91	0.470	
	RHY_DeltaV	-0.15	1.06	0.11	0.94	0.133	
	RHY_DeltaC	-0.14	1.06	0.10	0.95	0.137	
	RHY_VnPVI	-0.14	1.22	0.10	0.79	0.091	
	RHY_CrPVI	-0.12	1.08	0.09	0.94	0.105	
	RHY_VarcoV	-0.14	1.08	0.10	0.93	0.407	
	RHY_VarcoC	-0.15	1.11	0.11	0.90	0.120	
Lexical	LEX_ContDens	-0.17	1.09	0.12	0.91	0.003	*
	LEX_PoS_ADJ	-0.30	1.01	0.22	0.93	0.002	*
	LEX_PoS_ADV	-0.03	1.07	0.02	0.95	0.825	
	LEX_PoS_ART	0.04	1.04	-0.03	0.97	0.835	
	LEX_PoS_CONJ	-0.20	1.01	0.14	0.97	0.116	
	LEX_PoS_NOUN	-0.05	1.09	0.03	0.93	0.407	
	LEX_PoS_NUM	-0.00	1.01	0.00	1.00	0.989	
	LEX_PoS_PHRAS	0.19	1.42	-0.14	0.48	0.428	
	LEX_PoS_PREDET	0.04	1.25	-0.03	0.77	0.907	
	LEX_PoS_PREP	-0.14	1.03	0.10	0.97	0.124	
	LEX_PoS_PRON	0.11	1.08	-0.08	0.93	0.378	
	LEX_PoS_VERB	0.07	1.15	-0.05	0.88	0.082	
	LEX_RefRReal	-0.01	1.07	0.01	0.95	0.232	
	LEX_PDEIXIS	0.07	1.14	-0.05	0.88	0.407	
	LEX_SDEIXIS	-0.07	0.94	0.05	1.04	0.771	
	LEX_TDEIXIS	0.09	1.15	-0.06	0.88	0.525	
	LEX_RPRO	0.04	1.17	-0.03	0.86	0.736	
	LEX_NEGADV	0.15	1.19	-0.11	0.82	0.267	
	LEX_TTR	0.21	1.05	-0.15	0.93	0.040	*
	LEX_BrunetW	-0.27	1.09	0.20	0.88	0.028	*
	LEX_HonoreR	-0.00	1.13	0.00	0.90	0.536	
	LEX_ACTVRB	0.08	1.07	-0.06	0.94	0.307	
	LEX_DM_F	-0.08	1.08	0.06	0.94	0.491	
	LEX_IDEAD	-0.16	1.07	0.12	0.93	0.137	
	LEX_NW	-0.18	0.92	0.13	1.04	0.003	**
Syntactic	SYN_NPLENM	-0.15	1.00	0.11	0.99	0.417	
	SYN_NPLENSD	-0.14	1.10	0.10	0.91	0.267	
	SYN_GRAPHDISTM	-0.34	0.99	0.25	0.94	<0.001	***
	SYN_GRAPHDISTSD	0.04	1.07	-0.03	0.95	0.882	
	SYN_ISynCompl	-0.01	1.21	0.01	0.82	0.315	
	SYN_MAXDEPTHM	-0.26	0.79	0.19	1.09	0.002	**

(*continued*)

Table 9 (*Continued*)

Feature Type	Feature	MCI (PG)		Controls (CG)		KS	
		μ	σ	μ	σ	p-value	Signif.
	SYN_MAXDEPTHSD	-0.06	1.06	0.05	0.96	0.049	*
	SYN_SLENM	-0.31	0.83	0.23	1.06	<0.001	***
	SYN_SLENSD	-0.15	1.09	0.11	0.91	<0.001	***

Table 10
Means (μ) and standard deviations (σ) of the automatic classifiers results (macro-averaged F1-score) over 10 runs. In boldface the best result for each method/corpus/task combination.

Method (Corpus)	All Tasks		Task FIGURE		Task WORK		Task DREAM	
	μ	σ	μ	σ	μ	σ	μ	σ
RFC (AAC)	0.6887	0.0194	**0.6487**	0.0144	**0.6842**	0.0213	**0.6831**	0.0249
SVC (AAC)	**0.7045**	0.0185	0.5952	0.0380	0.6417	0.0302	0.6269	0.0197
RFC (MCC)	0.7030	0.0191	**0.6628**	0.0489	0.6828	0.0310	**0.7146**	0.0365
SVC (MCC)	**0.7445**	0.0164	0.6187	0.0322	**0.6856**	0.0196	0.6706	0.0326

Table 11
Previous works for MCI detection directly comparable to this study. F1 values marked with a * have been computed by us on the basis of the other values provided by the authors in their paper.

Reference	Language	ML Method	Best results
Vincze et al. (2016)	Hungarian	SVC	F1 = 68.9%
Asgari et al., 2017	English	SVC	Acc = 0.76; Sens = 0.53
			Spec = 0.88; F1=71.7%*
Tóth et al., 2018	Hungarian	RFC (Automatic)	F1 = 76.0%
		SVC (Manual)	F1 = 75.0%
Themistocleous et al., 2018	Swedish	DNN	F1 = 65.8%*
Gosztolya et al. (2019)	Hungarian	SVC	F1 = 78.3%
Fraser et al. (2019a)	Swedish	SVC	Acc = 0.69; Sens = 0.54
			Spec = 0.84; F1=68.3%*
Fraser et al., 2019	English	SVC	Acc = 0.63; Sens = 0.53
			Spec = 0.74; F1=62.8%*
	Swedish		Acc = 0.72; Sens = 0.77
			Spec = 0.67; F1=71.9%*
This study	Italian	SVC (Automatic)	F1 = 70.5%
		SVC (Manual)	F1 = 74.5%

- this is a clear indication of data scarcity when trying to devise automatic classifiers trained and evaluated using a limited amount of linguistic examples. Each subject produced a single recording for each linguistic task, thus the classifier for any single task is trained by using one-third of the data available for training the classifier working on all tasks. This is a typical behaviour for ML techniques when trained with scarce data;
- the three tasks have been properly designed to involve different cognitive skills: in addition to stimulating the semi-spontaneous verbal production of the subjects, allowing the subsequent linguistic analysis, they allow to evaluate the possible breakdown of the memory functions. As a matter of fact, in all three tasks it is essential to remember what you are saying (working memory), what you have already referred to (episodic memory) or what you are going to tell (prospective memory); moreover, the completion of the task requires the knowledge of the lexemes used, including semantic and lexical information (semantic memory), and the ability to recall, upon request, personal memories referring to a more or less remote past (autobiographical, recent and remote episodic memory). Obviously all the tasks involve all the memory systems, however the involvement has different "percentages". For example, FIGURE impacts a great deal on semantic memory, but does not involve autobiographical episodic memory. If the different tasks provide different contributions to describe the subject's state, then it does not sound so strange that the classifier using the complete "picture" working on all data is able to identify the different subjects with better performance.

For single tasks, RFCs exhibit the best performance in all combinations except one, suggesting that this ML technique is less influenced by data scarcity with respect to SVCs. On the contrary, SVCs are consistently better when applied to the whole datasets including all tasks.

Table 12
Means (μ) and standard deviations (σ) of the automatic classifiers results (macro-averaged F1-score) over 10 runs for the different feature families considering the SVC technique.

Corpus	Feature set	All Tasks	
		μ	σ
AAC	Acoustic	0.5972	0.0366
	Demographic	0.3888	0.0239
	Readability	0.3577	0.0273
	Rhythmic	0.5228	0.0355
	Lexical	0.4960	0.0628
	Syntactic	0.6014	0.0319
	ALL	**0.7045**	0.0185
AAC	Significant	0.6662	0.0391
MCC	Acoustic	0.5847	0.0392
	Demographic	0.3888	0.0239
	Readability	0.4968	0.0456
	Rhythmic	0.5713	0.0555
	Lexical	0.4570	0.0437
	Syntactic	0.5990	0.0607
	ALL	**0.7445**	0.0164
MCC	Significant	0.7126	0.0150

It seems relevant to note the small difference in performance exhibited by the fully automatic procedure for deriving all the features in an automatic way (70.5%) with respect to the classifier results when trained on manually checked data (74.5%). Considering the goal of the project discussed in the previous sections, this seems the most promising result, suggesting that we can automatically extract all the features needed to classify a new subject in a reliable way from raw interview recordings.

In spite of the growing body of research on the topic, a good practice for reporting research has not yet been established for this specific task. As already noticed by Gosztolya et al. (2019) the choice of evaluation metric is not a clear-cut issue for this task: thus, a strict comparison among systems is not easy to draw. Moreover, previous works have already exploited different machine learning techniques and train-test configurations, and experiments have been conducted on a limited number of languages (i.e. mainly English, Swedish, Spanish, French, Hungarian), making it difficult to contrast different cohorts. Results have been reported variously in the existing literature. ROC (Receiver Operating Characteristics) curve is quite common among clinicians as a performance measurement for classification problem at various thresholds settings; it plots sensitivity versus specificity across a range of values for the power to predict a dichotomous outcome. Some relevant paper in the field reports the ”Area Under the Curve” (AUC) and the ”Equal Error Rate” (EER), which are two metrics derived from the ROC: they correspond, respectively, to the area subtended by the curve and the point where the false positive rate and the false negative rate are equal, namely the intersection of the ROC curve with the straight line of 45 degrees.

Some classical Information Retrieval metrics are also widespread in the literature: “accuracy” (i.e. the number of correct predicted samples over the total number of samples), “precision” (i.e. the fraction of relevant samples among the retrieved samples) and “recall” (i.e. the fraction of the total amount of relevant samples actually retrieved). These last two scores are usually aggregated in the “F-measure” (or “F1-score”), which corresponds to their harmonic mean. However, accuracy alone is often reported; this can be misleading, especially in the presence of class-imbalance.

Given this unfolding scenario, in order to discuss our results, we conduct a brief review of related works exploiting linguistic biomarker for MCI automatic detection in various languages. Published studies that included the following have been considered for eligibility: i) employment of machine learning classifier ii) confronting healthy subject and MCI patients for screening purpose iii) published in English iv) no restrictions on the language spoken by the enrolled subjects. Instead, studies have been excluded if: i) cohorts included early or frankly dementia patients ii) results have been reported as AUC, EER and accuracy alone.

Table 11 summarises the selected papers: we inserted only the best result, when building a classifier for distinguishing controls from MCI subjects, and the associated ML system for brevity. Most of the cited works in Table 11 performed a lot of experiments with different settings: in particular, the very best results claimed by the respective authors were based on systems trained by using only the significant (or the N-most significant) features. As discussed before, in our opinion this is not the most suitable way to evaluate system performance, because this procedure introduces a bias, artificially inflating the final results. As a matter of fact, the selection of the statistically significant features has been done by examining the entire dataset of subjects’ productions, considering also the data that has been used in the subsequent system test phase. Even if we did the statistical analysis of the linguistic features we considered in this study, all the experiments presented here used the whole set of available features, avoiding the introduction of any bias in the evaluation phase. For this reason in Table 11 we report only the best results obtained in the other studies without any feature selection, allowing for a fair comparison with our setting.

Making specific reference to the cited studies, from Gosztolya et al. (2019) we extracted the results on 2 classes without feature selection; in Fraser et al. (2019a) and Fraser et al., 2019 we took into consideration the results on two-class discrimination exploiting linguistic features alone and not considering non-ecological settings like eye tracking or the reading aloud task; the study from Tóth et al., 2018 follows similar procedure as the one we are presenting, but, even if it sounds very surprising that the fully automatic system is able to achieve better results then the algorithm trained on manually checked data, it is very interesting seeing such good performance for a completely automatic procedure. With regard to the work of Themistocleous et al., 2018 we averaged the F1-score reported in the paper for each fold to obtain a single value.

Given these premises and considerations, we can say that our results on Italian are in line, or better, with the state-of-the-art for other languages presenting, on average, a F1-score around 75%. In order to compare our results with the works from Fraser et al. (2019a), Fraser et al., 2019 and Asgari et al., 2017 we computed the macro-averaged F1 score from the data on their papers (accuracy, sensitivity and specificity) thus we can safely compare our F1-score with their results.

As a general observation, it is interesting to note that none of the most recent studies in this field makes use of DNN, confirming our observation that for small datasets it is better to use traditional ML techniques. The only exception making use of DNN, the paper from Themistocleous et al., 2018, did not produce results near the state-of-the-art reinforcing our choice not to use neural networks for this kind of task.

In order to disentangle the real contribution of the different feature families, we tested a set of classifier by using a single group of features. We made these experiments by using SVCs, the technique exhibiting the best general results in the experiments discussed before. Table 12 shows the results for any group of features considered in this study as well as the results obtained using only the significant features from Table 9. Acoustic and Syntactic features confirm their good ability to distinguish MCI subjects from controls, as already evidenced by statistical significance, but it is interesting to note that, despite the fact that no rhythmic feature resulted significant, all of them, when taken together, are able to bring some contribution to the classification process. Actually, 5 over 7 rhythmic features are not so far from significance threshold. By looking at the results provided by the classifiers based only on the significant features, well below from the best ones obtained using all features, we can observe that, even if the contribution of significant features is certainly relevant to sustain classifier performance, also non significant features are able to bring useful contributions to the system improving the performance by various points. In our opinion, this could be another argument in favour of not to build classifiers using only significant features and, instead, use the full set of available features, avoiding any kind of bias in the experiments and taking advantage of the contributions, even partial, of less significant, or non significant but in any case relevant, features.

As a final comment, it should be pointed out that a complete language-specific profiling of pathological verbal productions by means of computational techniques, despite its time-consuming nature, is an essential preliminary step for the implementation of a valid, reliable dementia screening instrument. From a linguistic point of view, typological differences (e.g. at the acoustical, morphological, syntactic and lexical level) might strongly limit the extension of the results, hindering the spread of similar tools in different geographical areas. Most of the studies focused on English, just as it is supposed to be. Therefore, the relevance of a wide range of variants should be tested from time to time, especially on less-described languages. In this respect, we hope that the number of studies on this topic will continue to grow.

4. Conclusion

This study presented a novel system for the detection of Mild Cognitive Impairment conditions in Italian, by examining subjects' productions during three spontaneous speech tasks.

We created a complex set of algorithms for the automatic extraction of several linguistic features, from the acoustic, rhythmic, lexical, syntactic and readability domains; then we build some ML classifiers, with the aim of discriminating healthy controls from MCI subjects. This system was able to perform the task exhibiting a macro-averaged F1-score around 75%, the state-of-the-art performance for more studied languages; this is a very encouraging result for our project. Examining the results obtained in this pilot study, we can reliably claim that the former dream of building tools helping for a massive screening of cognitive impairments directly by practitioners can be a true reality in the next few years.

As far as we know, this is the first study on Italian language examining a large set of linguistic features for building automatic classifiers identifying Mild Cognitive Impairment from healthy controls.

CRediT authorship contribution statement

Laura Calzà: Conceptualization, Supervision, Funding acquisition, Project administration. **Gloria Gagliardi:** Data curation, Validation, Writing - original draft. **Rema Rossini Favretti:** Conceptualization. **Fabio Tamburini:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Writing - original draft.

Acknowledgments

This work was supported by the OPLON project (Opportunities for active and healthy LONgevity, Smart Cities, Ministero Università e Ricerca, SCN_00176). The study was approved by the Ethical Committee of Azienda Ospedaliera Reggio Emilia (n. 148 2013/0013438). Given the particular kind of data employed for this study and the restrictions on them from the Italian legislation, unfortunately we cannot make the datasets publicly available. Daniela Beltrami and Enrico Ghidoni are gratefully acknowledged for subjects selection and interview recordings.[9]

[9] As far as academic requirements are concerned, the abstract, 1, 2.1 and 4 have been authored by GG; FT takes official responsibility for 2, 2.3, 2.4, 2.5, 3 and 4. All authors read and gave final approval for submission.

References

Altmann, L.J.P., Kempler, D., Andersen, E.S., 2001. Speech errors in Alzheimer's disease: reevaluating morphosyntactic preservation. Journal of Speech, Language, and Hearing Research 44, 1069–1082.
Arias-Vergara, T., Vasquez Correa, J.C., Orozco-Arroyave, J.R., Nöth, E., 2018. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun 101 (101), 11–25.
Asgari, M., Kaye, J., Dodge, H., 2017. Predicting Mild Cognitive Impairment from spontaneous spoken utterances. Alzheimer's & Dementia: Translational Research & Clinical Interventions 3 (2), 219–228.
Ash, S., McMillan, C., Gross, R.G., Cook, P., Morgan, B., Boller, A., Dreyfuss, M., Siderowf, A., Grossman, M., 2011. The organization of narrative discourse in Lewy body spectrum disorder. Brain & Language 119, 30–41.
Austin, J.L., 1962. How to do things with words. Clarendon Press, Oxford.
Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., Calz, L., 2018. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? Front Aging Neurosci 10, 369.
Benba, A., Jilbab, A., Hammouch, A., 2016. Voice assessments for detecting patients with Parkinson's diseases using PCA and NPCA. Int J Speech Technol 19 (4), 743–754.
Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5 (9/10), 341–345.
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, 1–21.
Bosco, C., Lombardo, V., Vassallo, D., Lesmo, L., 2000. Building a Treebank for Italian: a Data-driven Annotation Schema. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000). ELRA - European Language Resources Association.
Brunet, E., 1978. Le vocabulaire de Jean Giraudoux. structure et evolution. Slatkine, Geneve.
Bucks, R.S., Singh, S., Cuerden, J.M., Wilcock, G.K., 2000. Analysis of spontaneous, conversational speech in dementia of alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology 14 (1), 71–91.
Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, E., Marcello, N., Rossini-Favretti, R., Tamburini, F., 2015. Should we screen for cognitive decline and dementia? Maturitas 82 (1), 28–35.
Cantos-Gòmez, P., 2009. Featuring linguistic decline in Alzheimer's disease: A corpus-based approach. In: Mahlberg, M., Gonzàlez Dìaz, V., Smith, C. (Eds.), Proceedings of the Corpus Linguistics Conference 2009 (CL2009).
Carlomagno, S., Santoro, A., Menditti, A., Pandolfi, M., Marini, A., 2005. Referential communication in Alzheimer's type dementia. Cortex 41 (4), 520–534.
Catricalà, E., Della Rosa, P., Plebani, V., Perani, D., Garrard, P., Stefano Cappa, F., 2015. Semantic feature degradation and naming performance. evidence from neurodegenerative disorders. Brain & Language 147, 58–65.
Chapman, S.B., Zientz, J., Weiner, M.F., Rosenberg, R.N., Frawley, W.H., Burns, M.H., 2002. Discourse changes in early Alzheimer disease, Mild Cognitive Impairment, and normal aging. Alzheimer Disease & Associated Disorders 16 (3), 177–186.
Chinaei, H., Currie, L.C., Danks, A., Lin, H., Mehta, T., Rudzicz, F., 2017. Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. Computational Linguistics 43 (2), 377–406.
Ciurli, P., Marangolo, P., Basso, A., 1996. Esame del linguaggio - II. Giunti Organizzazioni Speciali, Firenze.
Conti, S., Bonazzi, S., Laiacona, M., Masina, M., Vanelli Coralli, M., 2015. Montreal Cognitive Assessment (MoCA) - italian version: regression based norms and equivalent scores. Neurological Science 26, 209–214.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20 (3), 273–297.
Cresti, E., Moneglia, M., 2018. The Illocutionary Basis of Information Structure: the Language into Act Theory (L-AcT). Chapter 13. The illocutionary basis of information structure, Information Structure in Lesser-described Languages: Studies in prosody and syntax, pp. 360–402.
Cruttenden, A., 1986. Intonation. Cambridge University Press, Cambridge.
de Lira, J.O., Ortiz, K.Z., Carvalho Campanha, A., Ferreira Bertolucci, P.H., Cianciarullo Minett, T.S., 2011. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. International Psychogeriatrics 23 (3), 404–412.
De Mauro, T., 1999. GRADIT: Grande dizionario italiano dell'uso. UTET, Torino.
De Mauro, T., 2000. Il dizionario della lingua italiana. Paravia.
Delbeuck, X., Debachy, B., Pasquier, F., Moroni, C., 2013. Action and noun fluency testing to distinguish between Alzheimer's disease and dementia with Lewy bodies. J Clin Exp Neuropsychol 35 (3), 1–10.
Dell'Orletta, F., Montemagni, S., Venturi, G., 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pp. 73–83.Edinburgh, Scotland, UK
Delwo, V., 2006. Rhythm and Speech Rate: A Variation Coefficient for deltaC. In: Karnowski, P., Szigeti, I. (Eds.), Language and Language-Processing. Peter Lang, Frankfurt am Main, pp. 231–241.
Drummond, C., Coutinho, G., Fonseca, R.P., Assunção, N., Teldeschi, A., de Oliveira-Souza, R., Moll, J., Tovar-Moll, F., Mattos, P., 2015. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. Front Aging Neurosci 7 (96).
Epelbaum, S., Genthon, R., Cavedo, E., Habert, M.O., Lamari, F., Gagliardi, G., Lista, S., Teichmann, M., Bakardjian, H., Hampel, H., Dubois, B., 2017. Preclinical Alzheimer's disease: a systematic review of the cohorts underlying the concept. Alzheimer's & dementia: the journal of the Alzheimer's Association 13 (4), 454–467.
Fergadiotis, G., Wright, H.H., 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology 25, 1414–1430.
Ferris, S.H., Farlow, M., 2013. Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. Clin Interv Aging 8, 1007–1014.
Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., Brambati, S.M., 2019. Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. Aphasiology.
Flach, P., 2012. Machine learning: The art and science of algorithms that make sense of data. Cambridge University Press, USA.
Folstein, M., Folstein, S.E., McHugh, P.R., 1975. Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Reserach 12, 189–198.
Forbes-McKay, K., Shanks, M.F., Venneri, A., 2013. Profiling spontaneous speech decline in Alzheimer's disease: a longitudinal study. Acta Neuropsychiatr 25 (6), 320–327.
Fraser, K.C., Lundholm Fors, K., Eckerström, M., Öhman, F., Kokkinakis, D., 2019. Predicting MCI status from multimodal language data using cascaded classifiers. Front Aging Neurosci 11, 205.
Fraser, K.C., Lundholm Fors, K., Kokkinakis, D., 2019. Multilingual word embeddings for the assessment of narrative speech in Mild Cognitive Impairment. Computer Speech & Language 53, 121–139.
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. Cortex 55, 43–60.
Fraser, K.C., Meltzer, J.A., Rudzicz, F., 2016. Linguistic features identify Alzheimer's disease in narrative speech. J. Alzheimers Dis. 49, 407–422.
Gagliardi, G., 2014. Validazione dell'ontologia dell'azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment. Università degli Studi di Firenze PhD Thesis.
Garrard, P., 2009. Cognitive archaeology: uses, methods, and results. J Neurolinguistics 22 (3), 250–265.
Garrard, P., Maloney, L.M., Hodges, J.R., Patterson, K., 2005. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128, 250–260.

Gorno-Tempini, M.L., Hillis, A., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J., Rohrer, J.D., Black, S.E., Boeve, B., Manes, F., Dronkers, N., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M., Grossman, M., 2011. Classification of primary progressive aphasia and its variants. Neurology 76, 1006–1014.
Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., Hoffmann, I., 2019. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Computer Speech & Language 53, 181–197.
Grabe, E., Low, E.L., 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. In: Gussenhoven, C., Warner, N. (Eds.), Papers in Laboratory Phonology 7. Mouton de Gruyter, Berlino, pp. 515–546.
Grossman, M., Gross, R., Moore, P., Dreyfuss, M., McMillan, C.T., Cook, P.A., Ash, S., Siderowf, A., 2012. Difficulty processing temporary syntactic ambiguities in Lewy body spectrum disorder. Brain & Language 120, 52–60.
Hirst, D., Di Cristo, A., 1998. Intonation systems: A survey of twenty languages. Cambridge University Press.
Hoffmann, I., Nemeth, D., Dye, C., Pàkàski, M., Irinyi, T., Kàlmàn, J., 2010. Temporal parameters of spontaneous speech in Alzheimer's disease. Int J Speech Lang Pathol 12 (1), 29–34.
Holmes, D.I., Singh, S., 1996. A stylometric analysis of conversational speech of aphasic patients. Literary and Linguistic Computing 11 (3), 133–140.
Honoré, A., 1979. Some simple measures of richness of vocabulary. Association of Literary and Linguistic Computing Bulletin 7, 172–177.
Jarrold, W.L., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J., 2014. Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech. In: Resnik, P., Resnik, R., Mitchell, M. (Eds.), Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. ACL - Association for Computational Linguistics, pp. 27–37.
Jiang, H., Hu, B., Li, Z., Yan, L., Wang, T., Liu, F., Kang, H., Li, X., 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Commun 90, 39–46.
Kemper, S., LaBarge, E., Ferraro, F., Kemper, S., Richard Cheung H, F., Cheung, H., Storandt, M., 1993. On the preservation of syntax in Alzheimer's disease. Arch. Neurol. 50 (1), 81–86.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6 (1), 10.
Leech, G., 2005. Adding Linguistic Annotation. In: Wynne, M. (Ed.), Developing Linguistic Corpora: a Guide to Good Practice. Oxbow Books, Oxford.
Lesmo, L., 2007. Il parser basato su regole del gruppo NLP dell'Università di Torino. Intelligenza Artificiale IV (2), 46–47.
Cuetos, F., Arango-Lasprilla, J.C., Uribe, C., Valencia, C., Lopera, F., 2007. Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease. J Int Neuropsychol Soc 13 (3), 433–439.
López-de-Ipiña, K., Alonso, J.-B., Travieso, C.M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., Martinez de Lizardui, U., 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors 13, 6730–6745.
López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J., Travieso, C., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., Beitia, B., 2015. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: a fractal dimension approach. Computer Speech & Language 30 (1), 43–60.
March, E.G., Wales, R., Pattison, P., 2006. The uses of nouns and deixis in discourse production in Alzheimer's disease. J Neurolinguistics 19, 311–340.
Matsuda Toledo, C., Aluisio, S.M., Borges dos Santos, L., Dozzi Brucki, S.M., Sturzeneker Trés, E., Okada de Oliveira, M., Lessa Mansur, L., 2018. Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10, 31–40.
Mazzeo, S., Padiglioni, S., Bagnoli, S., Bracco, L., Nacmias, B., Sorbi, S., Bessi, V., 2019. The dual role of cognitive reserve in subjective cognitive decline and Mild Cognitive Impairment: a 7-year follow-up study. J. Neurol. 266, 487–497.
Measso, G., Cavarzeran, F., Zappal, G., Lebowtz, B.D., Crock, T., Pirozzolo, F.J., Amaducci, L.A., Massari, D., Grigoletto, F., 1993. The Mini-Mental State Examination: normative study of an italian random sample. Dev Neuropsychol 9 (2), 77–95.
Meilán, J., Martínez-Sánchez, F., Carro, J., Carcavilla, N., Ivanova, O., 2018. Voice markers of lexical access in mild cognitive impairment and Alzheimer's disease. Curr Alzheimer Res 15, 111–119.
Mesulam, M., 2001. Primary progressive aphasia. Ann. Neurol. 49, 425–432.
Mesulam, M., 2003. Primary progressive aphasia: alanguage-based dementia. N top N. Engl. J. Med. 349 (16), 1535–1542.
Mortamais, M., Ash, J.A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Pose, C., Albala, B., Ropacki, M., Ritchie, C.W., Ritchie, K., 2017. Detecting cognitive changes in preclinical Alzheimer's disease: a review of its feasibility. Alzheimer's & Dementia 13, 468–492.
Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H., 2005. The Montreal Cognitive Assessment, MoCA: abrief screening tool for Mild Cognitive Impairment. J Am Geriatr Soc 53 (4), 695–699.
Nucci, M., Mapelli, D., Mondini, S., 2012. The cognitive reserve questionnaire (CRIq): a new instrument for measuring the cognitive reserve. Aging Clin Exp Res 24 (3), 218–226.
Patterson, C., 2018. The World Alzheimer Report 2018. Technical Report. Alzheimer's Disease International.
Petersen, R.C., 2011. Clinical practice. Mild Cognitive Impairment. New Engl. J. Med. 364 (23), 2227–2234.
Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, M., 2015. The World Alzheimer Report 2015. Technical Report. Alzheimer's Disease International.
Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292.
Ripich, D.N., Carpenter, B.D., Ziol, E.W., 2000. Conversational cohesion patterns in men and women with Alzheimer's disease: a longitudinal study. International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists 35 (1), 49–64.
Ritchie, K., Ropacki, M., Albala, B., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Ritchie, C.W., 2017. Recommended cognitive outcomes in preclinical Alzheimer's disease: consensus statement from the European prevention of Alzheimer's dementia project. Alzheimer's & Dementia 13, 186–195.
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.A., 2011. Spoken language derived measures for detecting Mild Cognitive Impairment. IEEE Transactions on Audio Speech, and Language Processing 19 (7), 2081–2090.
Sajjadi, S.A., Patterson, K., Tomek, M., Nestor, P.J., 2012. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. Aphasiology 1–20.
dos Santos, L.B., Corrêa Jr., E.A., Oliveira Jr, O.N., Amancio, D.R., Mansur, L., Aluísio, S.M., 2017. Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1284–1296.
Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., Tsolaki, M., 2013. Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia. INTERSPEECH 2013, 14th annual conference of the International Speech Communication Association. ISCA, pp. 1692–1696.
Singh, S., Bucks, R.S., Cuerden, J.M., 2001. An evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. Aphasiology 15 (6), 571–583.
Sirts, K., Piguet, O., Johnson, M., 2017. Idea density for predicting Alzheimer's disease from transcribed speech. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada, pp. 322–332.
Snowdon, D.A., 2003. Healthy aging and dementia: findings from the nun study. Ann. Intern. Med. 139 (5), 450–454.
Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R., Markesbery, W.R., 1996. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. The journal of the American Medical Association 275, 528–532.
Sorianello, P., 2006. L'allineamento tonale: alcune riflessioni. In: Savy, R., Crocco, C. (Eds.), Analisi Prosodica: teorie, modelli e sistemi di annotazione. Atti del secondo convegno AISV - Associazione Italiana di Scienze della Voce. EDK editore, Torriana (RN).

Stasak, B., Epps, J., Goecke, R., 2019. Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis. Speech Commun 115, 1–14.
Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., Pakaski, M., 2015. Speaking in Alzheimer's disease, is that an early sign? importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci 7, 195.
Szmrecsányi, B.M., 2004. On Operationalizing Syntactic Complexity. In: Purnelle, G., Fairon, C., Dister, A. (Eds.), Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Presses Universitaires de Louvain, Louvain-la-Neuve, pp. 1031–1038.
Sztahó, D., Vicsi, K., 2016. Estimating the Severity of Parkinson's Disease Using Voiced Ratio and Nonlinear Parameters. In: Král, P., Martín-Vide, C. (Eds.), Statistical Language and Speech Processing. Springer International Publishing, Cham, pp. 96–107.
Taler, V., Phillips, N.A., 2008. Language performance in Alzheimer's disease and Mild Cognitive Impairment: a comparative review. J Clin Exp Neuropsychol 30 (5), 501–556.
Themistocleous, C., Eckerström, M., Kokkinakis, D., 2018. Identification of Mild Cognitive Impairment from speech in swedish using deep sequential neural networks. Front Neurol 9, 975.
Tin, K.H., 1998. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20 (8), 832–844.
Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákáski, M., Kálmán, J., 2018. A speech recognition-based solution for the automatic detection of Mild Cognitive Impairment from spontaneous speech. Curr Alzheimer Res 15, 1–10.
Tsoi, K.K., Chan, J., Hirai, H., Wong, S.Y., Kwok, T., 2015. Cognitive tests to detect dementia: asystematic review and meta-analysis. JAMA Internal Medicine Sign In Individual Sign In Sign inCreate an Account 175 (9), 1450–1458.
Upadhya, S.S., Cheeran, A.N., Nirmal, J.H., 2019. Discriminating parkinson diseased and healthy people using modified mfcc filter bank approach. Int J Speech Technol 22 (4), 1021–1029.
Velayudhan, L., Ryu, S.-H., Raczek, M., Philpot, M., Lindesay, J., Critchfield, M., Livingston, G., 2014. Review of brief cognitive tests for patients with suspected dementia. International Psychogeriatrics 26 (8), 1247–1262.
Vigorelli, P., 2004. La conversazione possibile con il malato Alzheimer. Franco Angeli, Milano.
Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., Kálmán, J., 2016. Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 181–187.
Wang, T., Lian, C., Pan, J., Yan, Q., Zhu, F., Ng, M.L., Wang, L., Yan, N., 2019. Towards the Speech Features of Mild Cognitive Impairment: Universal Evidence from Structured and Unstructured Connected Speech of Chinese. In: Proc. Interspeech 2019, pp. 3880–3884.
Yancheva, M., Fraser, K., Rudzicz, F., 2015. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT).
Yancheva, M., Rudzicz, F., 2016. Vector-space topic models for detecting Alzheimer's disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 2337–2346.
Yu, H., Mak, M., 2011. Comparison of voice activity detectors for interview speech in nist speaker recognition evaluation. In: Proc. of Interspeech'11, pp. 2353–2356.