Received: 7 January 2022 | Accepted: 27 February 2023

DOI: 10.1111/1460-6984.12870

International Journal of Language & Communication Disorders

REVIEW

# Natural language processing techniques for studying language in pathological ageing: A scoping review

**Gloria Gagliardi**

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

**Correspondence**
Gloria Gagliardi, Department of Classical Philology and Italian Studies, University of Bologna, Via Zamboni, 32, I-40126 Bologna, Italy.
Email: gloria.gagliardi@unibo.it

**Funding information**

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors

## Abstract

**Background:** In the past few years there has been a growing interest in the employment of verbal productions as *digital biomarkers*, namely objective, quantifiable behavioural data that can be collected and measured by means of digital devices, allowing for a low-cost pathology detection, classification and monitoring. Numerous research papers have been published on the automatic detection of subtle verbal alteration, starting from written texts, raw speech recordings and transcripts, and such linguistic analysis has been singled out as a cost-effective method for diagnosing dementia and other medical conditions common among elderly patients (e.g., cognitive dysfunctions associated with metabolic disorders, dysarthria).

**Aims:** To provide a critical appraisal and synthesis of evidence concerning the application of natural language processing (NLP) techniques for clinical purposes in the geriatric population. In particular, we discuss the state of the art on studying language in healthy and pathological ageing, focusing on the latest research efforts to build non-intrusive language-based tools for the early identification of cognitive frailty due to dementia. We also discuss some challenges and open problems raised by this approach.

**Methods & Procedures:** We performed a scoping review to examine emerging evidence about this novel domain. Potentially relevant studies published up to November 2021 were identified from the databases of MEDLINE, Cochrane and Web of Science. We also browsed the proceedings of leading international conferences (e.g., ACL, COLING, Interspeech, LREC) from 2017 to 2021, and checked the reference lists of relevant studies and reviews.

**Main Contribution:** The paper provides an introductory, but complete, overview of the application of NLP techniques for studying language disruption due to dementia. We also suggest that this technique can be fruitfully applied to other medical conditions (e.g., cognitive dysfunctions associated with dysarthria, cerebrovascular disease and mood disorders).

International Journal of Language & Communication Disorders



**Conclusions & Implications:** Despite several critical points need to be addressed by the scientific community, a growing body of empirical evidence shows that NLP techniques can represent a promising tool for studying language changes in pathological aging, with a high potential to lead a significant shift in clinical practice.

**KEYWORDS**
dementia, digital linguistic biomarkers, natural language processing, pathological ageing

## WHAT THIS PAPER ADDS

*What is already known on this subject*

- Speech and languages abilities change due to non-pathological neurocognitive ageing and neurodegenerative processes. These subtle verbal modifications can be measured through NLP techniques and used as biomarkers for screening/diagnostic purposes in the geriatric population (i.e., digital linguistic biomarkers—DLBs).

*What this paper adds to existing knowledge*

- The review shows that DLBs can represent a promising clinical tool, with a high potential to spark a major shift to dementia assessment in the elderly. Some challenges and open problems are also discussed.

*What are the potential or actual clinical implications of this work?*

- This methodological review represents a starting point for clinicians approaching the DLB research field for studying language in healthy and pathological ageing. It summarizes the state of the art and future research directions of this novel approach.

## INTRODUCTION

Natural language processing (NLP)—that is, the interdisciplinary field that aims to get computers to perform tasks involving human language (Jurafsky & Martin, 2008: 1)—is gaining popularity among the medical community. The applications are varied, ranging from research to diagnostics and direct patient care (Locke et al., 2021).

Notably, during the past few years there has been a growing interest in the employment of verbal productions as *digital biomarkers*, namely objective, quantifiable behavioural data that can be collected and measured by means of digital devices, allowing for a low-cost pathology detection, classification and monitoring (Gagliardi et al., 2021: 1). In particular, NLP techniques are increasingly used for clinical purposes in the geriatric population. Numerous research papers have been published on the automatic detection of subtle verbal alteration, starting from raw speech recordings and transcripts, and such linguistic analysis has been identified as a cost-effective method for diagnosing cognitive deterioration due to dementia. Moreover, in the present day, this novel approach is spreading to the identification of other clinical or psychiatric conditions in the elderly population, such as cognitive dysfunctions associated with metabolic disorders (e.g., type 2 diabetes mellitus; cf. Imre et al., 2019), depression (De Souza et al., 2021) and dysarthria (e.g., due to Parkinson's disease; cf. Rahman et al., 2021). Actually, such a cross-disciplinary approach presents a promising answer to the clinical challenge of cognitive assessment by combining objectivity (and reproducibility) in the evaluation process with unintrusiveness, speed and low cost.

## BACKGROUND

Speech and language are valuable sources of clinical information on cognitive status because they change depending on psychological distress (e.g., depressive symptoms; cf. Bernard et al., 2016; Tølbøll, 2019) and decline in parallel with neurodegeneration. A broad body of scientific literature suggests that linguistic deficits can be found in several neurodegenerative diseases (Boschi et al., 2017; Rahul & Ponniah, 2019), especially in dementia of the Alzheimer type, where language disruptions are common both at the early stages and in the full-blown pathology (Szatloczki et al., 2015), in combination with episodic memory and visuospatial impairments. Additionally, changes in verbal competence are receiving special attention as early signs of cognitive decline, since they are frequently present even in the preclinical and prodromal phases of the disease (i.e., subjective memory complaints—SMCs; and mild cognitive impairment—MCI; cf. McCullough et al., 2019). Consequently, linguistic performances have been extensively examined in the neuropsychological domain, in both research and clinical contexts, concerning the diagnostic process (Taler & Phillips, 2008; Filiou et al., 2020).

However, the assessment of verbal skills through standardized psychometric instruments is time-consuming and prone to human bias: Current pen-and-pencil tools often lack sufficient evidence of reliability to dementia patients (Krein et al., 2019). This is in line with the low validity of common screening tests, such as the Mini-Mental State Examination (MMSE) (Folstein et al., 1975) in identifying subtle deficits in cognitive functioning (Mitchell, 2009; Creavin et al., 2016; Hwang et al., 2019; Arevalo-Rodriguez et al., 2021).

An early diagnosis is a pivotal challenge to the promotion of the optimal management of cognitive frailty—irrespective of the reason—at both the individual and societal levels. Timely risk identification and a prompt, customized intervention might reduce the psychological burden of patients and caregivers, enabling the implementation of preventive measures and appropriate treatment. Besides, it represents a key strategy to contain the economic impact on social welfare and healthcare systems (Calzà et al., 2015).

In this background, automatic speech and language analysis through NLP methods has progressively acquired prominence. As stated by several scholars (Beltrami et al., 2018; de la Fuente Garcia et al., 2020; Petti et al., 2020), the usage of digital linguistic biomarkers (DLBs) has many advantages compared with classical paper-and-pencil tests. First, their computation is completely non-intrusive, time-effective and inexpensive. As such analysis does not require extensive infrastructure or medical equipment or laboratories, gathering information is easy and quick (König et al., 2018). These characteristics make DLBs particularly suitable as a life-course assessment. Second, this methodology can be administered remotely since it is easy to integrate with the existing telemedicine solutions. As the COVID-19 outbreak has shown, telehealth is of the utmost importance during extreme events, to allow optimal service delivery to fragile populations—such as elderly people—while reducing the risk of direct person-to-person contact (Bertini et al., 2022; König et al., 2021a). Finally, NLP techniques provide offline and online measures of cognitive activities underlying language production which would otherwise be impossible to be extracted and quantified by human operators (Gagliardi et al., 2021).

### Rationale of the study

In this scoping paper, peer-reviewed studies on the application of NLP techniques to the analysis of language changes in ageing were identified, collated, compared and evaluated to provide a useful resource for researchers to inform best practice. In particular, our purpose is to summarize the state of the art on this topic—with a special focus on cognitive frailty detection—through the analysis of the following key issues:

- What classification task is performed? What is the specific goal thereof?
- What type of linguistic data are collected and analysed? How?
- What classification methods are usually applied? How are they evaluated?

For each of these questions, some challenges and open problems are discussed.

## METHODS

In the absence of formal guidelines on the design and reporting of scoping studies (such as the PRISMA statement for systematic reviews; cf. Moher et al., 2009) we followed the suggestions of Arksey and O'Malley (2005), Munn et al. (2018) and Mbuagbaw et al. (2020). Our goal was to map the shreds of evidence available in this emerging area, being as comprehensive as possible in identifying relevant primary studies but at the same time allowing the replicability of the search outcomes.

The investigations by Barragán Pulido et al. (2020), Clarke et al. (2021a), de la Fuente Garcia et al. (2020), Petti et al. (2020) and Martínez-Nicolás et al. (2021) represent a valuable starting point for this work, since they provide

a consistent critical review of the considerable amount of papers published on the topic.

To base our observations on a comprehensive knowledge of the topic, the literature search was conducted in the databases MEDLINE, Cochrane and Web of Science, using the following keywords:

(dementia OR Alzheimer OR mild cognitive impairment OR cognitive frailty) AND (language OR speech) AND (NLP OR detection OR identification).

All databases were accessed between 1 and 3 September 2021. An updated search was performed on 20 November 2021. Further, we screened the proceedings of the following major international conferences in the field from 2017 to 2021:

- Annual Meeting of the Association for Computational Linguistics (ACL) (55th–59th editions: Long papers, short papers, student research workshop, system demonstrations, workshops).
- European Chapter of the Association for Computational Linguistics (EACL) (15th and 16th editions: Long papers, short papers, student research workshop, software demonstrations).
- Asian Chapter of the Association for Computational Linguistics (AACL) (1st edition: Research papers, student research workshop, system demonstrations, workshops).
- North American Chapter of the Association for Computational Linguistics (NAACL) (2018, 2019 and 2021 editions: Research papers, student research workshop, demonstrations, workshops).
- International Conference on Recent Advances in NLP (RANLP) (2017, 2019 and 2021 editions).
- International Conference on Computational Linguistics (COLING) (27th and 28th editions: Research paper, system demonstrations, workshops).
- Interspeech (2017–21 editions).
- International Conference on Language Resources and Evaluation (LREC) (11th and 12th editions).
- Resources and Processing of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID workshop) (2nd and 3rd editions).

In addition, the reference lists of relevant studies were also checked, for further confirmation. We then applied some exclusion criteria: Only English-language, peer-reviewed articles were considered and pre-print and unpublished works were not taken into account. Moreover, we discarded the papers lacking a clear description of the enrolled cohort. In the end, we based our observations on a sample of 179 papers.

## NLP TECHNIQUES FOR THE ANALYSIS OF LANGUAGE CHANGES DUE TO DEMENTIA: THE STATE OF THE ART

As mentioned in the Introduction, computational analysis of language changes due to ageing is performed through the estimation of DLBs. In this review, we will focus our attention on dementia assessment—and especially, among many proteins misfolding diseases, on Alzheimer's disease early identification—which has always been one of the most explored research areas in this field. However, this method can be seamlessly extended to a variety of clinical conditions, including but not limited to dysarthria, cerebrovascular disease and mood disorders (Gagliardi et al., 2021).

While the scientific literature documents a wide range of approaches to tackle the problem, it is possible to pinpoint at least some common steps:

- Collection (and eventual annotation) of verbal production written/uttered by patients and (matched) healthy controls.
- Automatic extraction of quantitative linguistic features and the optional preliminary testing of their discriminative powers.
- Application and validation of classification algorithms on these data.

Most of the time, studies adopt an observational retrospective case-control setting (Mann, 2003).

Machine learning (ML)—the subfield of artificial intelligence (AI) that is concerned with the question of how to construct computer programs that automatically improve with experience (Mitchell, 1997: xv) in order to perform an automated detection of meaningful patterns in data (Shalev-Shwartz & Ben-David, 2014: xv)—is pivotal to this research program. In short, in DLB research, speakers' data are usually annotated with labels for the target diagnosis (e.g., AD, Alzheimer's disease; MCI, mild cognitive impairment; SCI, subjective cognitive impairment; and HC/CON, healthy controls), established based on clinical evaluation (e.g., mainly neuropsychological testing, but including structural/functional brain imaging and fluid biomarkers). This information, known as the 'ground truth', is exploited to train the ML algorithm, so that it can 'learn' to discriminate the diagnostic classes in the training set. This procedure enables the system to predict the cognitive status of new, previously unseen speakers on a statistical basis, with some accuracy. Less frequently, an 'unsupervised' learning model is applied, in which case the algorithm is not provided with any pre-assigned labels for the training and must seek to structure unannotated data,

self-discovering the occurrence of any hidden patterns (Jo, 2021).

We attempt to answer the research questions posed in the Rationale section by outlining the various phases involved in this procedure, from task definition to communicative/cognitive deficit identification and prediction, passing by cohort selection, data-gathering and linguistic feature extraction.

## The DLB research field: Overall goals and task definition

The goals of the DLB research field are twofold: (1) to detect communicative impairments due to cognitive frailty in a screening/diagnostic perspective; and (2) to monitor the progression of linguistic symptoms throughout the disease trajectory.

Most of the works are devoted to the first aim: They try to distinguish the verbal productions of patients with a certain degree of cognitive impairment (i.e., subclinical, preclinical or full-blown pathology) and stage of neurodegeneration (i.e., early, moderate, severe) from those of healthy controls. Generally, even when datasets include three or more different cohorts, most studies perform only pairwise comparisons: Dementia versus healthy ageing. However, this may be incorrect from a clinical application perspective, to create viable tools for real-life environments. In fact, this dominant pairwise perspective does not grasp the complexity of cognitive frailty assessment in geriatric settings (Panza et al., 2015).

Just a few papers are devoted to fine-grained dementia classification (e.g., targeting the underpinning protein misfolding diseases), all dealing with frontotemporal degeneration subtyping (e.g., Fraser et al., 2014; Garrard et al., 2014; Nevler et al., 2019; Themistocleous et al., 2018; 2021; Cho et al., 2020). Moreover, despite mood disorders—especially depressive symptoms and apathy—being very common in older adults, causing reversible deterioration of cognitive performances, researchers have dealt only occasionally with the assessment of behavioural noncognitive disturbances of patients (e.g., König et al., 2019, 2021b; Sumali et al., 2020; Villatoro-Tello et al., 2021).

Limited progress has been achieved regarding the second goal too—that is, to follow the trajectory of the syndrome. To date, cross-sectional paradigms are still the most popular across the community: Although some corpora do include longitudinal speech samples, researchers have not focused on predicting the conversion from subclinical or preclinical stage to full-blown dementia, except in isolated cases (cf. Clark et al., 2016; Weiner & Schultz, 2016).

## Patient enrollment, data collection and DLB computation

A fundamental step of the DLB experimental approach is the recruitment of cohorts and linguistic data collection. Concerning this, papers should be explicit on inclusion/exclusion criteria for patient enrolment, following international guidelines (cf. Jack et al., 2011; McKhann et al., 2011; Albert et al., 2011; Sperling et. al., 2011; Dubois et al., 2014), to guarantee comparability with other studies. To the extent possible, the cohorts should be balanced—at least considering sex, age and education—since conclusions drawn from skewed corpora are subject to bias, especially in small datasets. Moreover, the statistical confirmation of cohort balance should be reported (as in Beltrami et al., 2018). Unfortunately, as observed by de la Fuente Garcia et al. (2020), not all research projects are in line with these basic design criteria.

DLBs can be detected both from written and oral texts, but the former approach is much less common (e.g., Toledo et al., 2014; Aramaki et al., 2016; Rentoumi et al., 2017). On the latter, different 'speaking styles' have been exploited: Read speech, repetition and (semi-)spontaneous speech. Moreover, researchers tested either telephone (e.g., Tröger et al., 2018; Yu et al., 2018) or on-site face-to-face recording set-ups (e.g., in clinical settings or during natural conversations). Occasionally, avatars and virtual agents were used to trigger verbal productions (e.g., Tanaka et al., 2017; Ujiro et al., 2018; Mirheidari et al., 2019; O'Malley et al., 2021). A popular elicitation strategy is the recording of verbal productions during standardized neuropsychological assessment: In this domain, the most widely used tasks are 'verbal fluency' (phonemic or semantic) and picture description (e.g., the Cookie Theft picture from the Boston Diagnostic Aphasia Examination; cf. Goodglass et al., 1984). Conversely, (semi-)spontaneous speech is usually triggered by interviewing the patients, engaging in conversation with them or asking them to recall something (e.g., a narrative plot, an event, a day or a dream). As the reader can observe, there is a high variability of study methods and settings in the scientific literature. Consequently, in our opinion, it is not possible to offer any advice on the more effective elicitation task at the moment. A deeper comparative analysis would be desirable, expanding the findings of Clarke et al. (2021b), Yamada et al. (2021) and Ivanova et al. (2022).

DLBs can be extracted directly from the audio files or the text/transcript. Transcription can be made manually or through automatic speech recognition (ASR) algorithms. The former strategy has reduced usability in the real-life context, due to its time-consuming and mistake-prone nature. However, it is also true that open-source ASR

tools are not currently reliable for pathological speech automatic recognition.

Considering the DLB computation, there is a remarkable variability among the studies regarding extraction methods. Most researchers adopt their own algorithm for feature extraction (e.g., Calzà et al., 2021; Gagliardi & Tamburini, 2022), but the situation is set to change over the next year, thanks to the increasing number of open-source tools and standardized feature sets (e.g., for acoustic indices, OpenSmile; Eyben et al., 2010). A striking number of features have been employed in the literature as proxy measures of cognitive disorder due to dementia. These can be classified into three main groups:

1. *Speech-based features* comprise DLBs directly extracted from audio samples. They convey both linguistic and paralinguistic information (i.e., vocal cues expressing emotions, irony, etc.). They can be further split into:
   a. *Acoustical* (e.g., López-de-Ipiña et al., 2015; 2018; Haider et al., 2020):
      - *Prosodic features*, concerning temporal properties of the speech (e.g., pause rate, phonation rate, speech rate, articulation rate) and fundamental frequency ($F_0$).
      - *Loudness and energy*.
      - *Spectral features*, such as formant trajectories (i.e., F1, F2 and F3), mel frequency cepstral coefficient (MFCC) and spectral centroid.
      - *Vocal quality*, such as jitter, shimmer and harmonic-to-noise ratio (HNR).
      - *ASR-related features*, such as disfluencies, repetitions, filled pauses and fractal dimension.
   b. *Rhythmic*, that is, variability of the syllabic intervals (cf. Martínez-Sánchez et al., 2016; Meilán et al., 2020; Calzà et al., 2021).
2. *Text-based features* consist of DLBs computed on written texts or transcripts. They gauge a wide range of linguistic dimensions, at multiple levels:
   a. *Lexicals* are DLBs which probe vocabulary richness (e.g., type-token ratio, Brunét's index, Honoré's statistics), the 'density' of verbal productions (e.g., content density, idea density), the rate of part of speech (e.g., nouns, verbs, adjectives, pronouns) or the incidence of specific lexical–semantic categories (cf. LIWC (Linguistic Inquiry and Word Count) ; Tausczik & Pennebaker, 2010).
   b. *Syntactical DLBs* measure the complexity of sentence structure on constituency- or dependency-based parse trees (cf. Roark et al., 2011; Lundholm Fors et al., 2018).
   c. *Semantic DLBs* explore the meaning of the texts (e.g., through matrix decomposition methods such as Latent semantic analysis (LSA) and Principal Component Analysis (PCA) , embeddings, topic modelling and sentiment analysis).
   d. *Pragmatic DLBs* quantify the usage of deictics and the coherence of the text.
3. *Extra-linguistic/multimodal features* such as gaze (e.g., Fraser et al., 2019a), smile (e.g., Tanaka et al., 2017) and gait (e.g., Shinkawa et al., 2019) are collected through wearable devices or sensors.

See Voleti et al. (2010: 284), de la Fuente Garcia et al. (2020: 1552) and Petti et al. (2020: 1791) for a detailed overview of the indices and their presumed discrimination power in dementia research. However, it is critical to highlight that the actual occurrence of specific linguistic cues in relation to different age-related cognitive profiles is under debate (Gagliardi & Tamburini, 2022). Language alterations brought on by dementia have been mostly reported in the lexical, syntactic and pragmatic domains. In a clinical perspective, they can be easily explained by the typical localization of brain atrophies (e.g., medial temporal lobe and hippocampus for AD) and linked to other cognitive alterations (e.g., episodic and semantic memory, executive functions). As a result, the computational analysis of these verbal competencies can be considered a reliable, well-established source of DLBs. However, it should be emphasized that algorithms usually display even higher performance on acoustical data (i.e., segmental and prosodic features of the speech, probably linked to anatomically driven phenomena). Unfortunately, the nature and the clinical relevance of these voice abnormalities remain largely unexplored. A larger body of evidence is needed to shed light on this point.

To conclude this section, we would like to emphasize that large linguistic corpora are essential for this research domain. However, their collection, filing, sharing and dissemination raises several ethical and legal issues that affect their full accessibility to the scientific community (e.g., pathological speech recordings and DLBs pertain to 'special category of personal data' according to the European Union's (EU) General Data Protection Regulation (GDPR), which imposes strict privacy rules). Several initiatives, such as the DELAD project (Lee et al., 2021), are currently in progress to overcome these problems. Nonetheless, currently, data scarcity represents one of the main limitations for the generalizability of the findings and translation of this technique into clinical practice.

## Classification through ML methods: Adopted algorithms and their evaluation

A large variety of algorithms have been applied to the task, ranging from 'conventional' supervised classifiers to deep

learning methods (see Aggarwal, 2019; and Jo, 2021, for a comprehensive picture of the field). The choice mostly depends on the dataset size.

The conventional supervised algorithm includes naïve Bayes classifiers, *k*-nearest neighbour, logistic regression, support vector machine (SVM), random forest (RF) or algorithms that produce interpretable outputs such as decision trees, applied singly or in combination. Since researchers usually deal with small datasets, a smaller number of studies have used artificial neural networks and deep learning methods, such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), recurrent neural networks (RNNs), multilayer perceptrons (MLPs) and autoencoders. See de la Fuente Garcia et al. (2020: 30–41, supplementary material) for a detailed survey. Very recently, some research groups adopted a transfer learning-based approach and language models (Yang et al., 2020), such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), with encouraging results (Haulcy & Glass, 2020; Balagopalan et al., 2021; Guo et al., 2021; Liu et al., 2021; Roshanzamir et al., 2021; Saltz et al., 2021). The more promising aspect of this novel methodology is the possibility of eliminating the time-consuming step of feature engineering and mitigating the need for a big dataset. However, along with other deep learning models, it raises an issue known as the 'black box' problem: Unlike feature-based approaches, these systems are 'opaque', that is, it is difficult to 'look inside' to explain their behaviour (Zednik, 2021: 1). In other words, it is not clear what information (i.e., set of features) in the input dataset determines the results. This point is very critical in a clinical setting because physicians should always be able to explain how a diagnosis was posed (Dashwood et al., 2021).

Building reliable ML classifiers is a challenging task. A decisive point is avoiding 'overfitting', that is, the model's inability to generalize to unseen data. In short, the algorithm performs well on the training set, but its results drop significantly when tested on unfamiliar data, because of the limited size of the dataset, its low quality or the task difficulty.

In this sense, the main criticalities arise from the validation of ML systems and their evaluation (including reporting strategies). On the first matter, following de la Fuente Garcia et al. (2020), the most established practice for testing the classifier is 'cross-validation' (CV) which consists of randomly splitting the dataset into equal 'folds' (i.e., segments), using different portions to train and test the model on subsequent iteration. According to the authors, CV is probably the best choice in this context of data scarcity, since conventional hold-out strategies would involve running the test on only a few observations. However, most studies have used the CV as an evaluation technique, but with a preliminary feature filtering (e.g., statistical comparison among the group by inferential tests or application of feature selection/dimension reduction methods, such as information gain, principal component analysis and minimum redundancy maximum relevance). In our opinion (Calzà et al., 2021), the aforementioned operation should be avoided because it artificially inflates the final performance feeding the algorithm with features picked beforehand, from the whole dataset (without dividing training and test sets).

Regarding the evaluation procedures, despite the growing volume of research on DLBs, a good practice for reporting the performances of ML systems has not yet been established across the NLP community. The choice of the evaluation metric is not a clear-cut issue for this task (Gosztolya et al., 2019; Calzà et al., 2021). Results are sometimes reported by providing the receiver operating characteristics (ROC) curve, which plots sensitivity versus specificity across a range of values for the power to predict a dichotomous outcome or the area under the curve (AUC) and the equal error rate (EER). These two metrics descend 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, that is, the intersection of the ROC curve with the straight line of 45°. Several classical information retrieval metrics are also widely used: 'accuracy' (i.e., the number of correctly 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 retrieved). These last two scores are usually combined in the 'F-measure' (or 'F1-score'), which corresponds to their harmonic mean. However, many studies have reported accuracy alone and this can be misleading, especially in the case of a class imbalance (Calzà et al., 2021).

The heterogeneity of the reporting strategy makes the results hardly comparable. This situation, combined with the extremely rapid development of the field, discourages us from proposing a comparison among systems' performances (which exceed 90% accuracy in AD detection but are significantly lower with MCI, around 75–80%). In this respect, Petti et al. (2020), Martínez-Nicolás et al. (2021) and especially de la Fuente Garcia et al. (2020) provide an extensive and updated overview.

## DISCUSSION AND CONCLUSIONS

In this work, we reviewed the current approaches to language description and evaluation through NLP techniques in the elderly. In particular, we discussed the state of the art on studying verbal productions in healthy

and pathological ageing and the latest research efforts to build non-intrusive language-based tools for the early identification of cognitive frailty due to dementia.

Summing up, a huge body of empirical evidence shows that automatic analysis of speech and language can represent a promising tool, with a high potential to spark a major shift to clinical practice. However, several concerns need to be addressed.

Unfortunately, the quality and quantity of information are currently insufficient to offer recommendations on the selection of tasks, features and algorithms. To date, as already observed by de la Fuente Garcia et al. (2020), it is unfair to compare on an equal footing the accuracy of ML methods, since their conclusions are drawn from data triggered by various elicitation techniques under different recording conditions, and the linguistic parameters are generated in a non-standardized way (i.e., with extraction procedures poorly and inconsistently described, avoiding replication studies). On the first point—that is, the lack of comparability due to datasets—it would be advisable to systematically compare the different approaches against balanced benchmarks adopting common metrics (Luz et al., 2021a), such as the shared task provided by ADReSS Challenges (Luz et al., 2020; 2021b) at InterSpeech conferences. However, this aspect is quite challenging for languages other than English, and especially for underrepresented languages. This is not a minor issue, since typological peculiarities (at the acoustical, morphological, syntactic and lexical levels) may strongly affect the comparability and transferability of results. To date, only a few papers adopted a multilingual approach, to identify DLBs that can generalize beyond a single corpus/language model (e.g., Fraser et al., 2019b; Gosztolya et al., 2021; Lindsay et al., 2021).

In our opinion, at the theoretical level, some general issues should also be considered. As previously stated, most of the works are devoted to the discrimination between subjects with a clear diagnosis of dementia and healthy controls, in a binary classification setting. However, this task is barely helpful from a clinical perspective (Calzà et al., 2021; Gagliardi & Tamburini, 2021): According to the World Health Organization (WHO) (2017), national programs should promote early diagnosis to foster healthy lifestyles and enable adequate treatment. Therefore, we would like to recommend focusing future research efforts on pre-clinical state detection, namely MCI or SCI.

Moreover, differential diagnosis, psychiatric comorbidities and longitudinal analyses should be a priority in future investigations. As a possible point of interest, a more granular description of MCI subtypes in the cohorts (i.e., amnestic/non-amnestic, simple-/multiple-domain) (Winblad et al., 2004) should be provided, since each of them may have different aetiologies (e.g., degenerative, vascular, psychiatric, medication side effects) and progressions (Petersen, 2004).

Nevertheless, *repeatability*—namely, the average variation over a short period in which disease-related decline is unlikely to be evident or manifested through the features (Stegmann et al., 2020:110)—should be considered seriously. It is crucial to distinguish normal variation in verbal productions (i.e., linked to stress, exhaustion, mood and motivation to perform the task) from disease-related linguistic changes, to build a reliable tool.

As evidenced by Pessin et al. (2017) and Rojas et al. (2020), the acoustical properties of the voice and the rhythmical aspects of the speech are not only altered by the disease progression, but also change over a lifetime, because of some physiological events that occur in senescence (i.e., changes in the speech mechanism affecting the respiratory, phonatory and supralaryngeal systems, such as the weakening of respiratory muscles, ossification of laryngeal cartilages and atrophy of facial, mastication and pharyngeal muscles). These voice disorders, known as 'presbiphonia', manifest perceptually to the point that listeners can judge a speaker's age fairly accurately from speech alone (Pettorino & Giannini, 2011). Analogous considerations can be made on the lexical–semantic and formal aspects of language (Peelle, 2019; Poulisse et al., 2019; Wulff et al., 2019). To address these hurdles and adequately capture the disease trajectory considering age effects and demographic factors, a more accurate stratification of cohorts is required. Potentially relevant features may include socio-economic status (SES) (APA, 2007) and cognitive reserve (Pettigrew & Soldan, 2019), namely the adaptability (i.e., efficiency, capacity, flexibility, etc.) of cognitive processes that helps to explain differential susceptibility of cognitive abilities or day-to-day function to brain ageing, pathology or insult (Stern et al., 2020: 1306). Further, dataset collection and system training should deal with diatopic variation, considering both cross- and intralinguistic differences (i.e., dialects and regional varieties spoken by the patients): Actually, this aspect represents one of the crucial problems for implementing a real large-scale tool (Barragán Pulido et al., 2020). We feel that these pieces of information will provide new insights into understanding the origin of the communicative impairment, its nature (e.g., phonatory, articulatory or cognitive), and will also support the discrimination of different aetiologies, boosting the sensitivity and specificity of the methodology.

The verbal alterations detected by ML classifiers are loosely ascribed to cognitive abnormalities, despite the lack of clarity about its origins. The changes are usually explained through a decrease in motor control, deterioration of mnemonic systems and deficits of executive functions, but to date, robust and trustworthy correlations are yet to be presented. It would hence be interest-

ing to investigate the neurobiological basis of linguistic biomarkers by linking deviant linguistic features with local brain atrophies or neural activity, combining (functional) neuroimaging studies and NLP methods.

Last but not the least, while most of the papers on the topic stress the potential of DLBs to outperform conventional screening approaches, only a few studies effectively implement the methodology in clinical research and medical practice (e.g., Martinez de Lizarduy et al., 2017; Martínez-Sánchez et al., 2018; König et al., 2021a; Rentoumi et al., 2021; Liang et al., 2022). To quote de la Fuente Garcia et al. (2020: 1548), despite progress in research, the small, inconsistent, single-laboratory and non-standardized nature of most studies has yielded results that are not robust enough to be aggregated and thereafter implemented toward those goals. This has resulted in gaps between research contexts, clinical potential and actual clinical applications of this new technology. DLBs, similar to other standard medical devices, will require specific developmental stages to be considered reliable and safe for use. A first, crucial step in this direction may be to move from the adoption of this approach in case-control designs to prospective investigations.

## ACKNOWLEDGEMENTS

Open Access Funding provided by Universita degli Studi di Bologna within the CRUI-CARE Agreement.

## CONFLICTS OF INTEREST STATEMENT

The author does not have any known competing financial interests that could have influenced the results reported in this paper.

## DATA AVAILABILITY STATEMENT

The full bibliography that supports the arguments presented in this study is available from the author, upon request.

## ORCID

*Gloria Gagliardi* https://orcid.org/0000-0001-5257-1540

## REFERENCES

Aggarwal, C.C. (2019) *Neural networks and deep learning: A textbook*. New York (NY): Springer Nature.

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., et al. (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging—Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 270–279.

APA—American Psychological Association, Task Force on Socioeconomic Status. (2007) *Report of the APA Task Force on Socioeconomic Status*. Washington, DC: American Psychological Association.

Aramaki, E., Shikata, S., Miyabe, M. & Kinoshita, A. (2016) Vocabulary size in speech may be an early indicator of Cognitive Impairment. *PLoS ONE*, 11(5), e0155195.

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

Arksey, H. & O'Malley, L. (2005) Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, 8(1), 19–32.

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. *Frontiers in Aging Neuroscience*, 13, 635945.

Pulido, M.L.B., Hernández, J.B.A., Ballester, M.Á.F., González, C.M.T., Mekyska, J. & Smékal, Z. (2020) Alzheimer's disease and automatic speech analysis: A review. *Expert Systems with Applications*, 150, 113213.

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? *Frontiers in Aging Neuroscience*, 10, 369.

Bernard, J.D., Baddeley, J.L., Rodriguez, B.F. & Burke, P.A. (2016) Depression, language, and affect: An examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology*, 35(3), 317–326.

Bertini, F., Allevi, D., Lutero, G., Montesi, D. & Calzà, L. (2022) Automatic speech classifier for Mild Cognitive Impairment and early Dementia. *ACM Transactions on Computing for Healthcare*, 3(1), 8.

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A. & Cappa, S.F. (2017) Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, 8, 269.

Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, G., Marcello, N., Rossini-Favretti, R., et al. (2015) Should we screen for cognitive decline and dementia? *Maturitas*, 82(1), 28–35.

Calzà, L., Gagliardi, G., Rossini Favretti, R. & Tamburini, F. (2021) Linguistic features and automatic classifiers for identifying Mild Cognitive Impairment and Dementia. *Computer Speech & Language*, 65, 101113.

Cho, S., Nevler, N., Shellikeri, S., Ash, S., Liberman, M. & Grossman, M. & (2020) Automatic classification of Primary Progressive Aphasia patients using lexical and acoustic features. In: Kokkinakis, D., Lundholm Fors, K., Themistocleous, C., Antonsson, M. & Eckerström, M. (Eds.) *Proceedings of the LREC 2020 Workshop on Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID-3)*. Paris: ELRA, 60–65.

Clark, D.G., McLaughlin, P.M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., et al. (2016) Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15(2), 113–122.

Clarke, N., Barrick, T.R. & Garrard, P. (2021b) A Comparison of connected speech tasks for detecting early Alzheimer's disease and mild cognitive impairment using Natural Language Processing and Machine Learning. *Frontiers in Computer Science*, 3, 634360.

Clarke, N., Foltz, P. & Garrard, P. (2021a) How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, 129, 446–463.

Creavin, S.T., Wisniewski, S., Noel-Storr, A.H., Trevelyan, C.M., Hampton, T., Rayment, D., et al. (2016) Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane Database of Systematic Reviews*, 2016(1), CD011145.

Dashwood, M., Churchhouse, G., Young, M. & Kuruvilla, T. (2021) Artificial intelligence as an aid to diagnosing dementia: An overview. *Progress in Neurology and Psychiatry*, 25(3), 42–47.

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. *Journal of Alzheimer's Disease*, 78(4), 1547–1574.

De Souza, D.D., Robin, J., Gumus, M. & Yeung, A. (2021) Natural Language Processing as an emerging tool to detect Late-Life Depression. *Frontiers in Psychiatry*, 12, 719125.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Volume 1 (Long and Short Papers)*. Stroudsburg (PA): ACL, 4171–4186.

Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., Blennow, K., et al. (2014) Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurology*, 13, 614–629.

Eyben, F., Wöllmer, M. & Schuller, B. (2010) OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In: Del Bimbo, A., Chang, S.F. & Smeulders, A. (Eds.) *MM '10: Proceedings of the 18th ACM* International Conference *on Multimedia*. New York (NY): ACM, 1459–1462.

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

Folstein, M.F., Folstein, S.E. & McHugh, P.R. (1975) Mini-mental state': A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189–198.

Fraser, K.C., Lundholm Fors, K., Eckerström, M., Öhman, F. & Kokkinakis, D. (2019a) Predicting MCI status from multimodal language data using cascaded classifiers. *Frontiers in Aging Neuroscience*, 11, 205.

Fraser, K.C., Lundholm Fors, K. & Kokkinakis, D. (2019b) 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., et al. (2014) Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, 55, 43–60.

Gagliardi, G., Kokkinakis, D. & Duñabeitia, J.A. (2021) Editorial: Digital Linguistic Biomarkers: Beyond Paper and Pencil Tests. *Frontiers in Psychology*, 12, 752238.

Gagliardi, G. & Tamburini, F. (2021) Linguistic biomarkers for the detection of Mild Cognitive Impairment. *Lingue e linguaggio*, XX(1), 3–31.

Gagliardi, G. & Tamburini, F. (2022) The Automatic Extraction of Linguistic Biomarkers as a Viable Solution for the Early Diagnosis of Mental Disorders. In: Calzolari, N. et al. (Eds.) *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*. Paris: ELRA, 5234–5242.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B. & Gorno-Tempini, M.L. (2014) Machine Learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, 55, 122–129.

Goodglass, H., Kaplan, E. & Weintraub, S. (1984) *Boston diagnostic aphasia examination*. Philadelphia (PA): Lea & Febiger.

Gosztolya, G., Balogh, R., Imre, N., López, E., Hoffmann, I., Vincze, V., et al. (2021) Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech. *Computer Speech & Language*, 69, 101215.

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.

Guo, Y., Li, C., Roan, C., Pakhomov, S. & Cohen, T. (2021) Crossing the 'Cookie Theft' corpus chasm: Applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Frontiers in Computational Science*, 3, 642517.

Haider, F., de la Fuente, S. & Luz, S. (2020) An assessment of paralinguistic acoustic features for detection of Alzheimer's Dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 272–281.

Haulcy, R. & Glass, J. (2021) Classifying Alzheimer's Disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11, 624137.

Hwang, A.B., Boes, S., Nyffeler, T. & Schuepfer, G. (2019) Validity of screening instruments for the detection of dementia and mild cognitive impairment in hospital inpatients: A systematic review of diagnostic accuracy studies. *PLoS ONE*, 14(7), e0219569.

Imre, N., Balogh, R., Gosztolya, G., Tóth, L., Várkonyi, T., Lengyel, C., et al. (2019) Automatic recognition of temporal speech features in type 2 diabetes mellitus with mild cognitive impairment. In: Baranyi, P., Esposito, E., Maldonato, M. & Vogel, C. (Eds.) *10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. Piscataway (NJ): IEEE, 27–28.

Ivanova, O. Meilán, J.J.G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T.E. & González, N.C. (2022) Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. *Computer Speech & Language*, 73, 101341.

Jack, C.R.J., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., et al. (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, 7(3), 257–262.

Jo, T. (2021) *Machine learning foundations: Supervised, unsupervised, and advanced learning*. Cham: Springer Nature.

Jurafsky, D. & Martin, J.H. (2008) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd edition, Hoboken (NJ): Prentice Hall.

König, A., Linz, N., Zeghari, R., Klinge, X., Tröger, J., Alexandersson, J., et al. (2019) Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *Journal of Alzheimer's Disease*, 69(4), 1183–1193.

König, A., Mallick, E., Tröger, J., Linz, N., Zeghari, R., Manera, V., et al. (2021b) Measuring neuropsychiatric symptoms in patients with early cognitive decline using speech analysis. *European Psychiatry*, 64(1), e64.

König, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., et al. (2018) Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, 15(2), 120–129.

König, A., Zeghari, R., Guerchouche, R., Duc Tran, M., Bremond, F., Linz, N., et al. (2021a) Remote cognitive assessment of older adults in rural areas by telemedicine and automatic speech and video analysis: Protocol for a cross-over feasibility study. *BMJ Open*, 11, e047083.

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

Lee, A., Bessell, N., van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., et al. (2021) The latest development of the DELAD project for sharing corpora of speech disorders. *Clinical Linguistics & Phonetics*, 36(2–3), 102–110. https://doi.org/10.1080/02699206.2021.1913514

Liang, X., Batsis, J.A., Zhu, X., Driesse, T.M., Roth, R.M., Kotz, D., et al. (2022) Evaluating voice-assistant commands for dementia detection. *Computer Speech and Language*, 72, 101297.

Lindsay, H., Tröger, J. & König, A. (2021) Language impairment in Alzheimer's Disease—Robust and explainable evidence for AD-related deterioration of spontaneous speech through multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13, 642033.

Liu, Z., Proctor, L., Collier, P.N. & Zhao, X. (2021) Automatic diagnosis and prediction of cognitive decline associated with Alzheimer's Dementia through spontaneous speech. In: Karim, H.A. (Ed.) *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. Piscataway (NJ): IEEE, 39–43

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A. & Kitchen, G.B. (2021) Natural Language Processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, 38, 4–9.

López-de-Ipiña, K., Alonso, J.B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., et al. (2015) On automatic diagnosis of Alzheimer's Disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7, 44–55.

Lopez-de-Ipiña, K., Martinez-de-Lizarduy, U., Calvo, P.M., Mekyska, J., Beitia, B., Barroso, N., et al. (2018) Advances on automatic speech analysis for early detection of Alzheimer Disease: A non-linear multi-task approach. *Current Alzheimer Research*, 15, 139–148.

Lundholm Fors, K., Fraser, K. & Kokkinakis, D. (2018) Automated syntactic analysis of language abilities in persons with Mild and Subjective Cognitive Impairment. *Studies in health technology and informatics*, 247, 705–709.

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2020) Alzheimer's Dementia recognition through spontaneous speech: The ADReSS Challenge. In: Meng, H., Xu, B. & Zheng T. (Eds.) *Proceedings of Interspeech 2020*. Grenoble: ISCA, 2172–2176.

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2021a) Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech. *Frontiers in Computer Science*, 3, 780169.

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2021b) Detecting cognitive decline using speech only: The ADReSS$_0$ Challenge. In: Heřmanský, H., Černocký, H. (Eds.) *Proceedings of Interspeech 2021*. Grenoble: ISCA, 3780–3784.

Mann, C.J. (2003) Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20, 54–60.

Martinez de Lizarduy, U., Calvo Salomón, P., Gómez Vilda, P., Ecay Torres, M. & López de Ipiña, K. (2017) ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, 4(1), e037.

Martínez-Nicolás, I., Llorente, T.E., Martínez-Sánchez, F. & Meilán, J.J.G. (2021) Ten years of research on automatic voice and speech analysis of people with Alzheimer's Disease and Mild Cognitive Impairment: A systematic review article. *Frontiers in Psychology*, 12, 620251.

Martínez-Sánchez, F., Meilán, J.J.G., Vera-Ferrandiz, J.A., Carro, J., Pujante-Valverde, I.M., Ivanova, O., et al. (2016) Speech rhythm alterations in Spanish-speaking individuals with Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, 24(4), 418–434.

Martínez-Sánchez, F., Meilán, J.J.G., Carro, J. & Ivanova, O. (2018) A prototype for the voice analysis diagnosis of Alzheimer's Disease. *Journal of Alzheimer's disease*, 64(2), 473–481.

Toledo, C.M., Cunha, A., Scarton, C. & Aluísio, S. (2014) Automatic classification of written descriptions by healthy adults: An overview of the application of Natural Language Processing and Machine Learning techniques to clinical discourse analysis. *Dementia & Neuropsychologia*, 8(3), 227–235.

Meilán, J.J.G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T.E. & Carro, J. (2020) Changes in the rhythm of speech difference between people with Nondegenerative Mild Cognitive Impairment and with preclinical Dementia. *Behavioural Neurology*, 2020, 4683573.

Mbuagbaw, L., Lawson, D.O., Puljak, L., Allison, D.B. & Thabane, L. (2020) A tutorial on methodological studies: The what, when, how and why. *BMC Medical Research Methodology*, 20, 226.

McCullough, K.C., Bayles, K.A. & Bouldin, E.D. (2019) Language performance of individuals at risk for Mild Cognitive Impairment. *Journal of Speech, Language, and Hearing Research*, 62(3), 706–722.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R. Jr, Kawas, C.H., et al. (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging—Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 263–269.

Mirheidari, B., Blackburn, D., O'Malley, R., Walker, T., Venneri, A., Reuber, M., et al. (2019) Computational cognitive assess-

ment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia. In: *ICASSP 2019 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway (NJ): IEEE, 2732–2736.

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. *Journal of Psychiatric Research*, 43(4), 411–431.

Mitchell, T.M. (1997) *Machine learning*. New York (NY): McGraw-Hill.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G, & for the PRISMA Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097.

Munn, Z., Peters, M.D.J., Stern, C., Tufanaru, C., McArthur, A. & Aromataris, E. (2018) Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*,18, 143.

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

O'Malley, R.P.D., Mirheidari, B., Harkness, K., Reuber, M., Venneri, A., Walker, T., et al. (2021) Fully automated cognitive screening tool based on assessment of speech and language. *Journal of Neurology, Neurosurgery, and Psychiatry*, 92(1), 12–15.

Panza, F., Seripa, D., Solfrizzi, V., Tortelli, R., Greco, A., Pilotto, A., et al. (2015) Targeting Cognitive Frailty: Clinical and Neurobiological Roadmap for a Single Complex Phenotype. *Journal of Alzheimer's Disease*, 47, 793–813.

Peelle, J. (2019) Language and Aging. In: de Zubicaray, G.I. & Schiller, N.O. (Eds.) *The Oxford Handbook of Neurolinguistics*. Oxford: Oxford University Press, 295–316.

Petersen, R.C. (2004) Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3), 183–194.

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

Pettigrew, C. & Soldan, A. (2019) Defining Cognitive Reserve and Implications for Cognitive Aging. *Current Neurology and Neuroscience Reports*, 19(1), 1.

Pettorino, M. & Giannini, A. (2011) The speaker's age: A perceptual study. In: Lee, W. & Zee, E. (Eds.) *Proceedings of the 17th International Congress of Phonetic Sciences—ICPhS*. Hong Kong: City University of Hong Kong, 1582–1585.

Pessin, A.B., Tavares, E.L., Gramuglia, A.C., de Carvalho, L.R. & Martins, R.H. (2017) Voice and ageing: Clinical, endoscopic and acoustic investigation. *Clinical Otolaryngology*, 42(2), 330–335.

Poulisse, C., Wheeldon, L. & Segaert, K. (2019) Evidence Against Preserved Syntactic Comprehension in Healthy Aging. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(12), 2290–2308.

Rahul, D.R. & Joseph, P.R. (2019) Language impairment in primary progressive aphasia and other neurodegenerative diseases. *Journal of Genetics*, 98, 95.

Rahul, D.R. & Ponniah, R.J. (2019) Language impairment in primary progressive aphasia and other neurodegenerative diseases. Journal of genetics, 98, 95.

Rahman, W., Lee, S., Islam, M.S., Antony, V.N., Ratnu, H., Ali, M.R., et al. (2021) Detecting Parkinson Disease using a web-based speech task: Observational study. *Journal of Medical Internet Research*, 23(10), e26305.

Rentoumi, V., Paliouras, G., Danasi, E., Arfani, D., Fragkopoulou, K., Varlokosta, S., et al (2017) Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In: Baranyi, P. (Ed.) *8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. Piscataway: IEEE, 33–38.

Rentoumi, V., Vassiliou, E., Demiraj, A., Pittaras, N., Mandalis, P., Alexandridou, M., et al. (2021) LANGaware: Leveraging Machine Learning on natural language for the early detection of neurodegenerative and psychiatric diseases. *Alzheimer's & Dementia*, 17(S11), e052520.

Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K. & Kaye, J. (2011) Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transaction on Audio, Speech, and Language Processing*, 19(7), 2081–2090.

Rojas, S., Kefalianos, E. & Vogel, A. (2020) How does our voice change as we age? A systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 Years of Age. *Journal of Speech, Language, and Hearing Research*, 63, 533–551.

Roshanzamir, A., Aghajan, H. & Soleymani Baghshah, M. (2021) Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21, 92.

Saltz, P., Lin, S.Y., Cheng, S.C. & Si, D. (2021) Dementia detection using transformer-based deep learning and Natural Language Processing models. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. Piscataway (NJ): IEEE, 509–510.

Shalev-Shwartz, S. & Ben-David, S. (2014) *Understanding machine learning: From theory to algorithms*. New York (NY): Cambridge University Press.

Shinkawa, K., Kosugi, A., Nishimura, M., Nemoto, M., Nemoto, K., Takeuchi, T., et al. (2019) Multimodal behavior analysis towards detecting mild cognitive impairment: Preliminary results on gait and speech. *Studies in Health Technology and Informatics*, 264, 343–347.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., et al. (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging—Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 280–292.

Stegmann, G.M., Hahn, S., Liss, J., Shefner, J., Rutkove, S.B., Kawabata, K., et al. (2020) Repeatability of commonly used speech and language features for clinical applications. *Digital Biomarkers*, 4, 109–122.

Stern, Y., Arenaza-Urquijo, E.M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., et al. (2020) Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. *Alzheimer's & Dementia*, 16(9), 1305–1311.

Sumali, B., Mitsukura, Y., Liang, K.C., Yoshimura, M., Kitazawa, M., Takamiya, A., et al. (2020) Speech quality feature analysis for classification of depression and dementia patients. *Sensors*, 20(12), 3599.

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. *Frontiers in Aging Neuroscience*, 7, 195.

Taler, V. & Phillips, N.A. (2008) Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5), 501–556.

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., et al. (2017) Detecting dementia through interactive computer avatars. *IEEE Journal of Translational Engineering in Health and Medicine*, 5, 2200111.

Tausczik, Y. & Pennebaker, J. (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

Themistocleous, C., Eckerström, M. & Kokkinakis, D. (2018) Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. *Frontiers in Neurology*, 9, 975.

Themistocleous, C., Ficek, B., Webster, K., denOuden, D.B., Hillis, A.E. & Tsapkini, K. (2021) Automatic Subtyping of Individuals with Primary Progressive Aphasia. *Journal of Alzheimer's Disease*, 79(3), 1185–1194.

Tølbøll, K.B. (2019) Linguistic features in depression: A meta-analysis. *Journal of Language Works—Sprogvidenskabeligt Studentertidsskrift*, 4(2), 39–59.

Tröger, J., Linz, N., König, A., Robert, P. & Alexandersson, J. (2018) Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. In: Minsky, N. & Osmani, V. (Eds.) *PervasiveHealth '18: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. New York (NY): ACM, 59–66.

Ujiro, T., Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T., et al. (2018) Detection of dementia from responses to atypical questions asked by embodied conversational agents. In: Yegnanarayana, B. (Ed.) *Proceedings of Interspeech 2018*. Grenoble: ISCA, 1691–1695.

Villatoro-Tello, E., Dubagunta, S.P., Fritsch, J., Ramírez-de-la-Rosa, G., Motlicek, P. & Magimai-Doss, M. (2021) Late fusion of the available lexicon and raw waveform-based acoustic modeling for Depression and Dementia recognition. In: Heřmanský, H. Černocký, H. (Eds.) *Proceedings of Interspeech 2021*. Grenoble: ISCA, 1927–1931

Voleti, R., Liss, J.M. & Berisha, V. (2010) A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 282–298.

Weiner, J. & Schultz, T. (2016) Detection of intra-personal development of Cognitive Impairment from conversational speech. *Speech Communication; 12. ITG Symposium*. Piscataway (NJ): IEEE, 1–5.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.O., et al. (2004) Mild cognitive impairment—Beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, 256(3), 240–246.

World Health Organization (WHO). (2017) *Global action plan on the public health response to dementia 2017–2025*. Geneva: WHO.

Wulff, D.U., De Deyne, S., Jones, M.N., Mata, R. & The Aging Lexicon Consortium. (2019) New perspectives on the aging lexicon. *Trends in Cognitive Sciences*, 23(8), 686–698.

Yamada, Y., Shinkawa, K., Kobayashi, M., Nishimura, M., Nemoto, M., Tsukada, E., et al. (2021) Tablet-based automatic assessment for early detection of Alzheimer's disease using speech responses to daily life questions. *Frontiers in Digital Health*, 3, 653904.

Yang, Q., Zhang, Y., Dai, W. & Pan, S.J. (2020) *Transfer learning*. Cambridge: Cambridge University Press.

Yu, B., Williamson, J.B., Mundt, J.C. & Quatieri, T.F. (2018) Speech-based automated cognitive impairment detection from remotely-collected cognitive test audio. *IEEE Access*, 6, 40494–40505.

Zednik, C. (2021) Solving the Black Box Problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34, 265–288.

**How to cite this article:** Gagliardi, G. (2024) Natural language processing techniques for studying language in pathological ageing: A scoping review. *International Journal of Language & Communication Disorders*, 59, 110–122. https://doi.org/10.1111/1460-6984.12870