Linguistic Markers of Subtle Cognitive Impairment in Connected Speech: A Systematic Review

Amélie B. Richard, Manon Lelandais, Karen T Reilly, Sophie Jacquin-Courtois

ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

Linguistic Markers of Subtle Cognitive Impairment in Connected Speech: A Systematic Review

Amélie B. Richard,[a,b] Manon Lelandais,[c] Karen T. Reilly,[b] and Sophie Jacquin-Courtois[b,d]

[a] Université Montpellier Paul-Valéry, Praxiling UMR 5267, France [b] Université de Lyon, Lyon Neuroscience Research Centre, France [c] Université Paris Cité, CLILLAC-ARP EA3967, France [d] Physical Medicine and Rehabilitation Department, Henry-Gabrielle Hospital, Hospices Civils de Lyon, France

ARTICLE INFO

Article History:
Received April 29, 2024
Revision received June 24, 2024
Accepted August 26, 2024

Editor-in-Chief: Cara E. Stepp
Editor: Susan L. Thibeault

https://doi.org/10.1044/2024_JSLHR-24-00274

ABSTRACT

Purpose: This systematic review covers the current stage of research on subtle cognitive impairment with connected speech. It aims at surveying the linguistic features in use to single out those that can best identify patients with mild neurocognitive disorders (mNCDs), whose cognitive changes remain underdiagnosed.
Method: We followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines and proposed a full definition of features for the analysis of speech features. Fifty-one studies met the inclusion criteria. Most of them focused on age-related progressive diseases and included fewer than 30 subjects.
Results: A total of 384 features labeled with 335 different names was retrieved, yielding various results in discriminating individuals with mNCDs from controls.
Conclusions: This finding highlights the need for harmonized labels to further investigate mNCDs with linguistic markers. We suggest two different ways of assessing a feature's reliability. We also point out potential methodological issues that remain to be resolved, along with recommendations for reproducible research in the field.

Neurocognitive disorders cover a large range of etiologies characterized by a decline in cognitive performance, either noticed by the individual or their relatives (American Psychiatric Association, 2022). Neurocognitive disorders are considered mild (mild neurocognitive disorders [mNCDs]) when the individual's independency is preserved. These include degenerative or acquired diseases, such as neurodegenerative pathologies (e.g., early-stage Alzheimer's disease [AD], Lewy body dementia), vascular pathologies (e.g., stroke), brain injuries (e.g., traumatic brain injury), and infectious or oncological contexts (e.g., COVID-19 sequalae, HIV, cancer-related cognitive impairment). mNCDs also refer to cognitive-behavioral syndromes such as mild cognitive impairment (MCI; Petersen, 2016) and subjective cognitive decline (Röhr et al., 2020), which are objective or subjective cognitive disorders for which no underlying neuropathology have been diagnosed. In some cases, these syndromes are the manifestation of the onset of a neurodegenerative disease.

mNCDs substantially affect an individual's quality of life (see, for instance, Henderson et al., 2019, regarding cancer-related cognitive impairment). Individuals with mNCDs report suffering from forgetfulness, high distractibility, trouble with multitasking, and difficulties with learning and language (e.g., Petersen, 2016, for MCI). Patients report substantial difficulties in their everyday lives, but current neuropsychological tests often partially reveal or fail to detect such subtle cognitive impairment. For instance, questionnaires designed to assess self-reported cancer-related cognitive impairment are weakly correlated with objective measures. This results in a discrepancy between the intensity of the cognitive difficulties reported by the patients and their scores, deemed in the range of norms (Areklett et al., 2024). One explanation is that these screening or diagnosis tests are typically conceived for the detection of more severe or advanced disorders, such as dementia for the Mini-Mental State Examination (Tangalos et al., 1996). Another explanation might be the lack of ecological validity of such tests, which test a single cognitive function in optimal conditions (Costa &

Correspondence to Amélie B. Richard: amelie.richard@inserm.fr. Amélie B. Richard and Manon Lelandais contributed equally to this work. Sophie Jacquin-Courtois and Karen T. Reilly contributed equally to this work. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Fardell, 2019), whereas mNCDs' symptoms express themselves in the patient's everyday life. Furthermore, the presence of confounding factors, such as fatigue, anxiety, or depression, make the assessment complex to interpret (Vos et al., 2020). Finally, an individual's expectations about their own cognitive performances may differ depending on their daily life functioning as well as their professional activity.

The absence of clear pathological results leads to underdiagnose individuals with mNCDs, leaving them with an invisible disability. New approaches to detect subtle cognitive changes are thus needed.

Investigating mNCDs With Markers in Connected Speech

The literature in health research shows a growing number of analyses focusing on speech and discourse. Studies have shown the interest of such linguistic analyses to track cognitive changes in an individual. These analyses can be conducted along with other objective (i.e., neuropsychological tests) and subjective (i.e., questionnaires) evaluations. Their noninvasive character, high sensitivity, ecological validity, and feasibility are some of the assets accounting for their increasing popularity (Bryant et al., 2017; Lanzi et al., 2023). Language is deemed susceptible to contain markers of subtle cognitive change because of its interaction with other cognitive functions such as memory, attentional processes, and executive functions. Such markers are likely to be present in all types of language production (e.g., reading, sentence repetition), but might be more easily detected in language in context, that is, connected speech. Connected speech is defined as stretches of speech whose production is organized according to phonological, syntactic, and pragmatic rules. There is, however, little consensus on the best linguistic markers for identifying individuals with mNCDs in the general population. This results in an extensive number of features with unclear labels and various measures, as highlighted by Boschi et al. (2017), hindering the replicability of such studies. There is thus a need for a well-described easy-to-use feature classification.

One scoping and two systematic reviews have been carried out on methods involving speech. These studies investigated AD in association with more severe cognitive impairment than mNCDs and MCI. They have found that both manual and automatic methods for speech recognition or analysis can distinguish individuals with AD from healthy controls or from individuals with MCI (de La Fuente Garcia et al., 2020; Filiou et al., 2020; Ivanova et al., 2023; Martínez-Nicolás et al., 2021). As pointed out by Filiou et al. (2020), the numerous methods and linguistic markers highlighted by these reviews raise the question of their clinical applicability. However, these reviews focused on two conditions, namely, AD and MCI. Regarding the use of features, Filiou et al. highlighted that semantic and fluency measures are the most likely to find a significant effect of group difference in early-stage AD or MCI. Since mNCDs cover a large range of underlying neuropathologies, we expect different patterns of linguistic markers depending on the etiology, as instigated in the work of Boschi et al. (2017), even though a direct mapping between linguistic features and pathologies/syndromes will not be the focus of this review. Furthermore, while linguistic markers for AD have been identified, little is known about those for subtle cognitive impairment.

Objectives of This Review

The aim of this review is to give an overview of the markers currently in use to detect or investigate mNCDs independently from the underlying pathology or cognitive-behavioral syndrome. This systematic review covers the current stage of research on subtle cognitive impairment with markers in connected speech by examining a large number of recent publications. It aims at surveying the long list of speech features in use to identify those that can best detect subtle cognitive impairment in patients with mNCDs, a population whose cognitive changes remain underdiagnosed, particularly in the case of acquired lesions. At the same time, it provides a current synthesis of the methods in use for the detection of subtle cognitive impairment with the help of linguistic markers.

Terminology Used on This Review

The terms *linguistic variable*, *linguistic feature*, and *linguistic marker* can be found in the literature either with different or interchangeable meanings. In this review, we use "feature" to refer to a specific speech unit or phenomenon, "variable" when the speech unit/phenomenon is an entry of a statistical or machine learning (ML) method, and "marker" when the speech unit/phenomenon can successfully identify the case population from controls. We also made a distinction between "feature" and "measure." "Measure" refers to the way the feature is analyzed (e.g., count, ratio). However, some linguistic features are inseparable from their measure such as "speech rate" and "articulation rate."

Organization of the Review

The introduction gave a brief description of mNCDs and spells out the main benefits and challenges of using connected speech for their detection. In the Method

section, we describe the article selection process, the data extraction and synthesis, and the study quality scale we used to assess the data. The Results section gives an overview of the characteristics of the studies selected and then focuses on trends in the extracted linguistic features and their power to discriminate individuals with mNCDs from a control population. The Discussion section addresses the critical topics and challenges in using features of connected speech for detecting mNCDs. We first focus on the high variability in the reported results in the studies selected and suggest two different ways of assessing a feature's reliability. We then point out potential methodological issues that remain to be resolved, along with recommendations for reproducible research in the field. In the Conclusion section, we raise questions that could prompt new methodological developments and offer a number of starting points for future studies in health and speech sciences. Finally, we make our data collection tables available on Open Science Framework (OSF), https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d, along with our search equations and study assessment scale.

Method

We searched the following electronic bibliographic databases: PubMed, ScienceDirect, Embase, Web of Science, Google Scholar. The protocol of this study was registered on 02/04/2023 in the PROSPERO database (ID number: CRD42023394729), and its preprint was uploaded to a non–peer review repository for protocols: https://doi.org/10.21203/rs.3.pex-2276/v1.

Eligibility Criteria

Only randomized controlled trials and case–control studies written and published in English or French as primary studies in journals with a peer-review process were included, even if the focus of the paper was on another language. Productions referred as "gray literature" were not searched.

Population

We selected studies that included a healthy control group and individuals with mNCDs. This incorporates individuals with a cognitive-behavioral syndrome (e.g., MCI, subjective cognitive decline). We also selected studies that included a pathological control group other than subjects with mNCDs since speech features may help discriminate one pathology from another one. Studies including individuals with major motor speech impairment (e.g., apraxia of speech) and/or individuals younger than 18 years or older than 75 years were excluded. Studies with fewer than 10 participants in the patient group and studies that included participants with moderate-to-severe or severe cognitive impairment were also excluded (i.e., scores below 21/30 for the Mini-Mental State Examination or 22/30 for the Montreal Cognitive Assessment). As for language, bilingualism was either an exclusion criterion stated by the reviewed studies or not mentioned.

Connected Speech Tasks and Neuropsychological Tests

Only studies using recorded connected speech tasks were included. Reading or writing were not considered as connected speech tasks. All studies had to include at least one standardized screening test or a full standardized battery for assessing cognition.

Search Strategy

We did our last search in July 2023 and included studies published between 2012 and July 2023. With the help of literature, we selected main keywords for the case population (McDonald, 2017) and connected speech tasks (Boschi et al., 2017). We added exclusion criteria such as "therapy" and "care" to avoid studies unrelated to diagnosis and "articulation" to avoid studies focusing on motor speech only. Using Boolean operators, our prototypical syntax was "case population" AND "connected speech task" AND "methods" NOT "exclusion criteria" (e.g., "mild cognitive impairment" AND "picture-based description" NOT "therapy"). OSF, Appendix A: https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d.

Selection Process

We followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement (Moher et al., 2009) and reported our process in a PRISMA flowchart using a modified version of the screening and selection procedure found in the work of Pati and Lorusso (2018) and Mateo (2020). Pati and Lorusso propose a step-by-step method to conduct a PRISMA systematic review with extensive details and examples. Mateo describes helpful, easy-to-use methodological tools implemented in Excel and Zotero.

We selected our articles for inclusion following a three-step procedure:

1. Prescreening: focused on publication language, publication type, duplicates, incorrect metadata—conducted by one author (A.B.R.).
2. Title and abstract screening: excluded studies that did not focus on connected speech, including studies that used naming tasks or verbal fluency tasks. The following exclusion criteria were applied: (a) studies involving individuals with major motor disorders (e.g., apraxia of speech), (b) studies focusing exclusively

on reading and writing tasks, (c) studies treating unrelated topics (e.g., qualitative studies in sociology focusing on discourse analysis).
This step was conducted independently by two authors (A.B.R. & M.L.). We ran an interrater agreement test, which reached a substantial 84% agreement (Cohen's κ = .68). Agreement for the remaining abstracts was achieved by both authors reviewing them again and discussing their eligibility.

3. Full-text selection: Articles were excluded using the following criteria: (a) *population*: no control group, mean age below 18 or above 75 years, fewer than 10 participants per case group, scores below normal cognition or mild disorders cutoffs at screening tests; (b) *tasks*: studies that did not include a connected speech task; (c) *analysis*: studies that did not explicitly analyze speech features; (d) *methods, statistics, and study design*: descriptive statistics only, meta-analyses; systematic reviews, case studies; qualitative studies; theoretical or methodological articles.

This step was conducted independently by two authors (A.B.R. and M.L.). An interrater agreement test was run and reached near perfect agreement (96% and Cohen's κ = .90).

Data Extraction Process

Data were classed into four categories: (a) population, (b) tasks, (c) tests, and (d) methods. The data in each category are described in Table 1. When studies included more than 10 linguistic variables, we chose to only report those markers discriminating patients with mNCDs from controls. This specifically applies to studies using ML methods where more than a hundred features were analyzed.

Table 1. Types of data extracted from the reviewed studies.

Category	Data extracted
Population	Number, age, language, for both case and control populations Pathology for case population
Connected speech task	Name of the connected speech task Type of elicited discourse according to the authors Type of elicited discourse according to our categorization
Tests	Name of the neuropsychological tests used Cognitive domains assessed according to the tests' authors
Methods	Name of the linguistic features extracted Feature extraction mode (manual vs. automatic) Statistical tests used for discriminating case population from control and their *p* value Machine learning models used

Due to the nature of the studies included in our review, we used a modified version of the National Institute of Health Quality Assessment of Case–Control Studies scale to evaluate the quality of each study and its risk of bias. We removed questions on concurrent controls and exposure/risks, which were not applicable, and we added five items on the connected speech tasks, the linguistic variables, and the statistical or ML methods used to compare case and control populations. Appendix B on OSF (https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d) describes the guidance and questions for assessing studies with our modified version of the National Institutes of Health (NIH) grid. Items 1–12 have been left unchanged, and Items 13–17 were added for the purpose of this systematic review. Two authors (A.B.R. and M.L.) independently conducted the risk of bias and quality assessment and then discussed to reach consensus on a paper's quality (low, fair, good). A study reached a good score when its method was deemed fully replicable, a fair score when it lacked descriptive information on the linguistic variables or when it mislabeled discourse type, and a low score when information was missing in two or more items of the scale, hindering its replicability.

Synthesis Methods

Feature Synthesis

The assessed cognitive functions were grouped based on the functions reported by the authors of the included papers and on the neuropsychological tests they reported using.

To facilitate the analysis of the linguistic features, they were clustered into specific, hierarchized categories by two language scientists (A.B.R. and M.L.) on the basis of similar characteristics (e.g., sound, meaning, word order). We, for instance, grouped all speech pause variables and then split them into silent pauses and filled pauses. When no detail was given on the characteristics of a pause (whether silent or filled), we left it into the larger, superior category labeled "pauses (not specified)." We finally grouped the categories into four large main linguistic domains: phonetics-prosody, lexicon-semantics, syntax-morphosyntax, and discourse. These four large domains are visible in Table 2 (first column). We chose a loose classification system because boundaries between smaller sized domains can be fuzzy, as some features might belong to more than one. Features that did not fit any specific categories either by being not readable by humans (which is sometimes the case with ML feature extraction) or by lacking information were ascribed to a linguistic domain. Our classification system is available on OSF: https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d.

Boschi et al. (2017) suggested classifying markers based on the descriptions reported by the study authors.

Table 2. Proposal for a harmonized terminology of linguistic markers.

Linguistic domain Category	Linguistic features	Definition	Examples
Phonetics-prosody Disfluencies	Silent pauses	Absence of intensity and *F*0, any interval where the amplitude is undistinguishable from that of the background noise above 200 ms (see for instance, Duez, 1982)	A: and this woman drives her kids to # school
	Filled pauses	Any type of filler particle that delays speech production: "A phonetic exponent which is segmentally structured, semantically empty, syntactically unconstrained, and does not show an interjectional function" (Belz, 2023).	A: and this woman drives her kids to uh school
	Breaks	Self-interruption from the speaker involving an incomplete delivery of a phonetic, syntactic, or morphological unit as a result of its abandonment Can lead to the resumption of a new unit as part of self-repair At the lexical level, includes deletions, substitutions, insertions, and articulation errors linked to missed phonological targets (Pallaud et al., 2019)	With self-repair: A: and this woman drives her daugh- her kids to school Without self-repair: A: and this woman drives her kids to #
	Repetitions	Duplication(s) of a phoneme, syllable, word, or phrase (Fox Tree, 1995)	A: and this woman drives her k- kids school
Speed of speech	Speech rate	Measure for speed and density per time unit, as the number of syllables produced per minute of speech (De Jong & Wempe, 2009) For studies relying on a fully automated data extraction, a robust definition can be found in He et al. (2023): "number of nuclei / (total nuclei duration + total internuclei duration)."	
	Articulation rate	Measure for speed and density per time unit, as the number of syllables produced per minute of speech, without silent pause or laughter time (Miller et al., 1984)	
Length of speech units	IPU duration	Interpausal units are speech segments separated by a 250-ms silent pause, with a minimum duration of 300 ms (Bigi & Priego-Valverde, 2022)	A: and this woman drives her kids to school (2390 ms) # (250 ms) I didn't know she had kids (1770 ms)
	Turn duration	Conversational turn-taking system. A speech turn (turn constructional unit) is a stretch of speech by one speaker during which other participants assume the role of listeners, until a point of "projected completion" (Sacks et al., 1974), called a transition relevance place (TRP), which can include a silent pause. A turn shift can occur (but does not have to) at a TRP. If no turn shift occurs at the TRP, the turn continues.	A: and this woman drives her kids to school (2390 ms) # B: Did you know she had kids?
	Speech duration	Total speech time duration. Whether this includes silent pauses and other vocal phenomena such as laughter has to be indicated.	A: and this woman drives her kids to school # I didn't know she had kids (4410 ms)

(*table continues*)

Table 2. *(Continued).*

Linguistic domain Category	Linguistic features	Definition	Examples
Lexicon-semantics Words	Lexical items	Refers to content (i.e., concepts, actions, beings or objects with clear mental representation) and open-class words (i.e., nouns, [non-auxiliary] verbs, adjectives, and [some] adverbs).	A: and this woman drives her kids to school
	Grammatical items	Refers to function and closed-class words No independent meaning on its own or no full-fledged semantic features Can be prepositions, pronouns, conjunctions, auxiliary verbs, or (some) adverbs	A: and this woman drives her kids to school
	Lexical richness	Also referred as lexical diversity and density, is used to assess the number produced by a speaker, showing the proficiency of a discourse Often measured with a type–token ratio (number of different words divided by the total number of words) or the Shannon's Entropy	
Syntax–morphosyntax	Clause	A subject and a predicate (**verb** + any complement) Sometimes labeled as sentence, utterance, phrase, or proposition	A: and this woman **drives** her kids to school.
	Clause completion	Refers to the ability for a participant to finish a clause in accordance with grammatical rules	
	Clause length	Number of lexical and grammatical items in a clause Whether this excludes or includes filler particles has to be specified.	
	Mean length of utterance (MLU)	In the studies we reviewed, the definition was the following: MLU is calculated by dividing the total number of words by the number of utterances (Galetto et al., 2013; Lowit et al., 2022).	
	Syntactic errors	Clauses with inconsistent syntax, grammar, or morphology (for instance using a noun in a verb slot or using an incorrect inflection for tense) Dependent on context of use and language variety (e.g., African American Vernacular English, General American, Southern American English)	A: and this woman have driven her kids to school.
	Item class	Also called word classes or lexical categories POS tagset: syntactic class of the item This list includes nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, articles, and interjections.	
	Syntactic complexity	No consensus yet This is what we propose based on the studies we reviewed. Number of items that are additional to the obligatory constituents in a clause Includes subordinators and coordinators Can also be number of complex sentences (made up of several clauses) as opposed to simple sentences (made up of one clause) This has to be made explicit.	

(*table continues*)

Table 2. *(Continued).*

Linguistic domain Category	Linguistic features	Definition	Examples
Discourse	Informativeness	Notion of relevance (Grice's maxim Be Relevant) Proportion of discourse segments on topic vs. off topic in the case of a thematic task (i.e., description or narrative). Main discourse structure vs. side discourse structure (Van Kuppevelt, 1995)	A: and this woman drives her kids to school # I didn't know she had kids # then she goes to the office
	Coherence	How different segments of discourse are made to relate to one another (Kehler, 2006; Labov, 2001; Wright et al., 2014)	Coherence: A: and this woman drives her kids to school # I didn't know she had kids # then she goes to the office Incoherence: A: and this woman drives her kids to school # I didn't know she had kids # we were only two at the office
	Cohesion	Consistent use of reference in speech Consistent use of cohesive devices such as anaphora (e.g., through pronouns and noun phrases), connectives, and simultaneity markers Can also include temporal cohesion (McNamara et al., 2010)	A: and this woman drives her kids to school # but she [another woman] doesn't know that
	Epistemic and pragmatic events	Epistemic events: use of any hedging expression (Schiffrin, 1987) A speaker's lexical reference to their own uncertainty or inability; use of modal auxiliaries; use of hedging discourse markers (see Cuenca & Crible, 2019, for the rhetorical and interpersonal domains in the classification of discourse markers) Pragmatic event: speech overlapping with another participant's turn; turn change (can be forced turn-taking as in interruptions); turn retention; use of conversational feedbacks (also called backchannels) Conversational feedbacks usually consist of brief signals produced by the interlocutor during the main speaker's speech and can be verbal (e.g., yes), vocal (e.g., mhm), and/or gestural (head movements, smiles). They are mandatory to update the shared knowledge (common ground) and promote the alignment between participants, which is necessary for mutual comprehension and success of the interaction (Boudin et al., 2021).	Epistemic events: A: and this woman drives her kids to school # I can't remember her name # then she goes to the office Pragmatic events: A: and this woman drives her kids to school # B: mhm # A: then she goes to the office
	Theory of mind (ToM)	Items showing the speaker is aware of others' different representations and mental states. Includes items that refer to the acknowledgment/consciousness of shared knowledge between participants, and those that manage it In the reviewed studies, ToM was measured with emotional tone and words semantically related to feelings (Baron-Cohen et al., 1985).	A: and this woman drives her kids to school (i.e., no needs to name the woman because the speaker knows that their co-speaker knows the woman in question)

Note. IPU = interpausal unit; POS = part-of-speech.

In total, they reported 120 markers and provided a definition for each of them. In addition to surveying the descriptions provided by the study authors we included, we searched the literature in language sciences for comprehensive feature definitions. We particularly focused on studies in speech production with experimental data if available, in an effort to build an interface between clinical studies and linguistic studies. This process resulted in 23 definitions with clear labels for markers. Table 2 gives full definitions as well as examples of the 23 labels we identified. We kept measurement types (e.g., count, duration, rate) apart from observed variables (i.e., linguistic feature) whenever possible to enhance precision and replicability for specialists in other fields.

Discourse Types

To compare the features depending on the speech task, we classified the task into four discourse types. Descriptive discourse gives details about a referent (i.e., a picture, an event) without any interpretation while narrative discourse refers to storytelling. Procedural discourse follows a logical structure to explain the steps of a process. Finally, semispontaneous discourse refers to speech that is elicited by context rather than a task, as is the case for everyday speech.

Results

The database search retrieved 815 records. A further seven records were manually identified for a total of 821 records. Duplicates and unrelated topic papers were discarded in a prescreening phase leaving 313 records. Of the 509 screened records, 312 were removed based on the title and abstract, leaving 197 records for full-text retrieval. Despite direct requests to the authors, four full texts were irretrievable. Out of the 193 full texts that were screened, we excluded 142 for various reasons, including inappropriate population, tasks, methods, and/or analyses. Figure 1 shows the full flowchart for the literature screening and selection process.

Figure 1. Literature screening and selection flowchart following Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines.

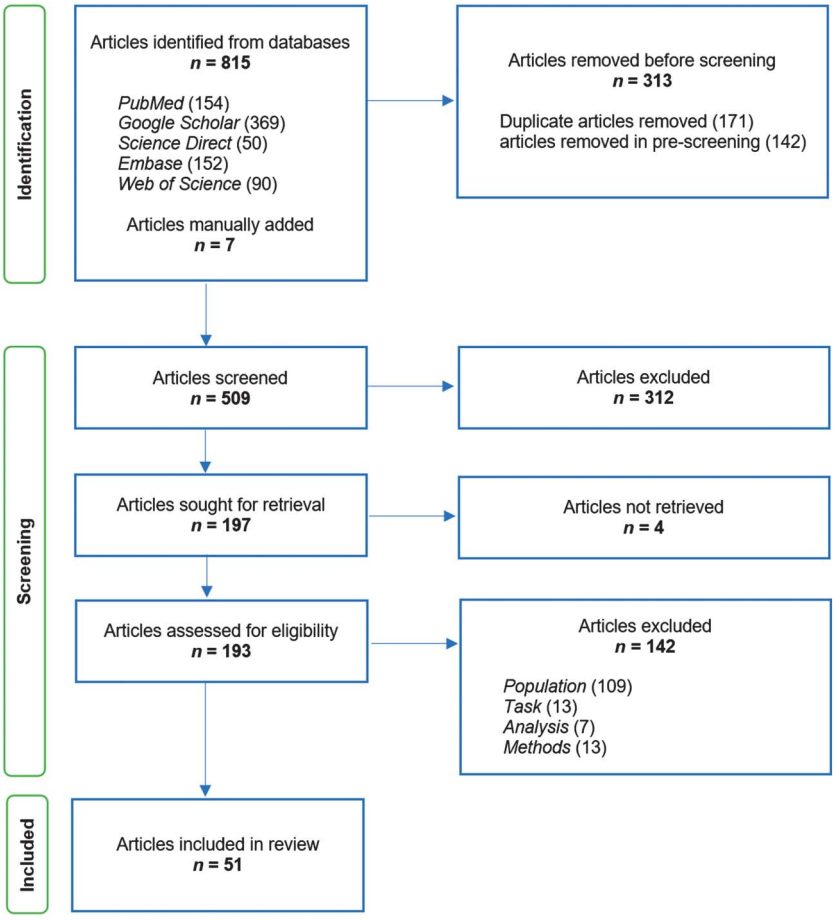
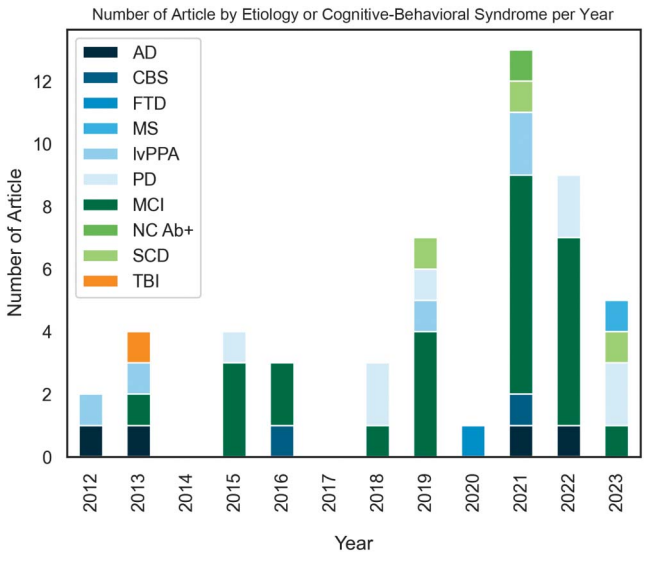


Figure 2. Number of studies per pathology per year. AD = Alzheimer's disease; CBS = cortico-basal syndrome; FTD = fronto temporal dementia; lvPPA = logopenic variant of primary progressive aphasia; MCI = mild cognitive impairment; MS = multiple sclerosis; NC Ab+ = normal cognitive with β-amyloid markers; PD = Parkinson's disease; SCI = subjective cognitive decline; TBI = traumatic brain injury; blue shades = neurodegenerative diseases; green shades = cognitive-behavioral syndromes; orange = nonneurodegenerative diseases.



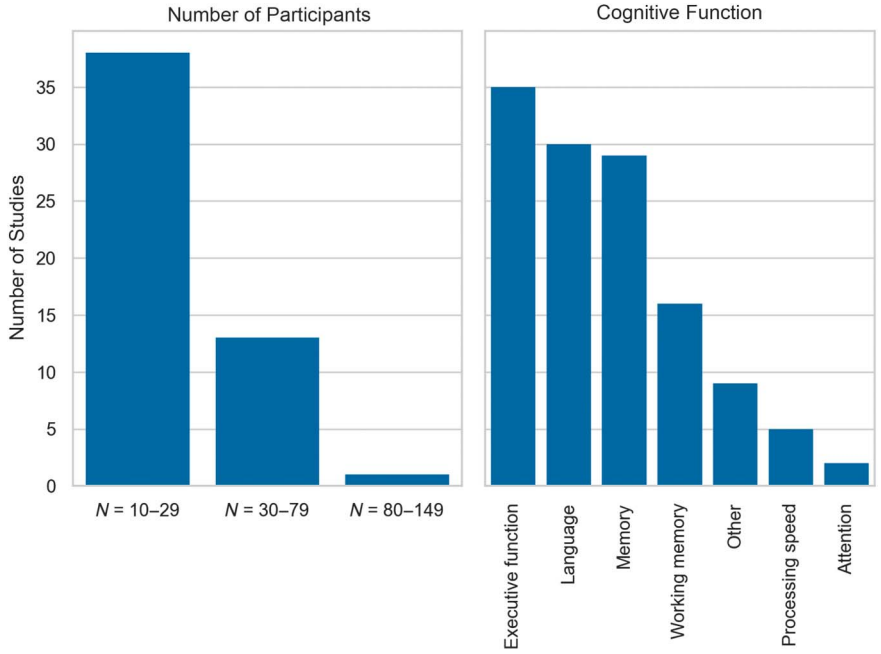
Population Overview

In this section, we summarize the main results. A comprehensive summary of the extracted information is available on OSF (https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d). The 51 studies included in this review came from 36 different research teams, and the majority were published within the last 3 years. Figure 2 shows the evolution in the number of studies per pathology from 2012 to July 2023. The number of studies was relatively stable between 2012 and 2018, and it more than doubled in the last 5 years, with the exception of 2020—due to the COVID-19 period, showing the growing interest for assessing subtle cognition impairment through connected speech analysis. The most represented etiologies and cognitive-behavioral syndromes across studies were neurodegenerative pathologies. Within these, studies mostly focused on MCI and early onset AD, followed by Parkinson's disease and the logopenic variant of primary progressive aphasia. Only five studies focused on nonneurodegenerative diseases (i.e., traumatic brain injury, subjective cognitive decline, normal cognition with β-amyloid markers).

Almost half of the studies included English-speaking participants (24), but a total of 14 languages were identified. Other frequent languages included Hungarian (six), French (three), Korean (three), and Swedish (three). One study did not report the participants' language.

Figure 3 shows the sample size (left) across studies as well as the assessed cognitive functions using neuropsychological tests (right). A majority of studies featured between 10 and 29 participants. Very few studies included more than 80 participants. As for cognitive evaluation, executive functions were very frequently assessed, as well as language and memory. However, attention was very rarely focused on.

Figure 3. Number of participants (left) and cognitive function assessed (right) across studies.



Linguistic Features

We now turn to linguistic features. We retrieved 384 features labeled with 335 different names, indicating that most of the features as reported verbatim were study specific, although they often referred to analogous features across studies.

Most of the studies did not use specifically developed software for the analysis of linguistic data. For studies that did use such software, Praat (Boersma & Weenink, 2024) was the most frequent tool. Features were equally extracted with manual methods and with semi- or fully automatized methods. Most of the studies did not use software for extracting the linguistic features (30/51 studies). When it was the case, Python scripts or ML models were used. The 16 studies using ML models for analyzing the features mostly used support vector machine (SVM) models. An interesting point is that the phonetics-prosody domain was the most investigated domain with ML models (11/16 studies).

Figures 4 and 5 show the relation between linguistic features and case populations as well as types of assessed discourse. Figure 4 shows the different linguistic features mapped out across the case pathologies or syndromes. Studies investigating MCI were by far the most frequent (25/51).

Figure 4. Linguistic feature frequency (count) according to case population. Lighter colors represent higher numbers of studies. Colors of the *y*-axis labels refer to the linguistic domains: blue = phonetics-prosody; red = lexical-semantics; green = morphosyntax–syntax; orange = discourse. *n* = number of studies.

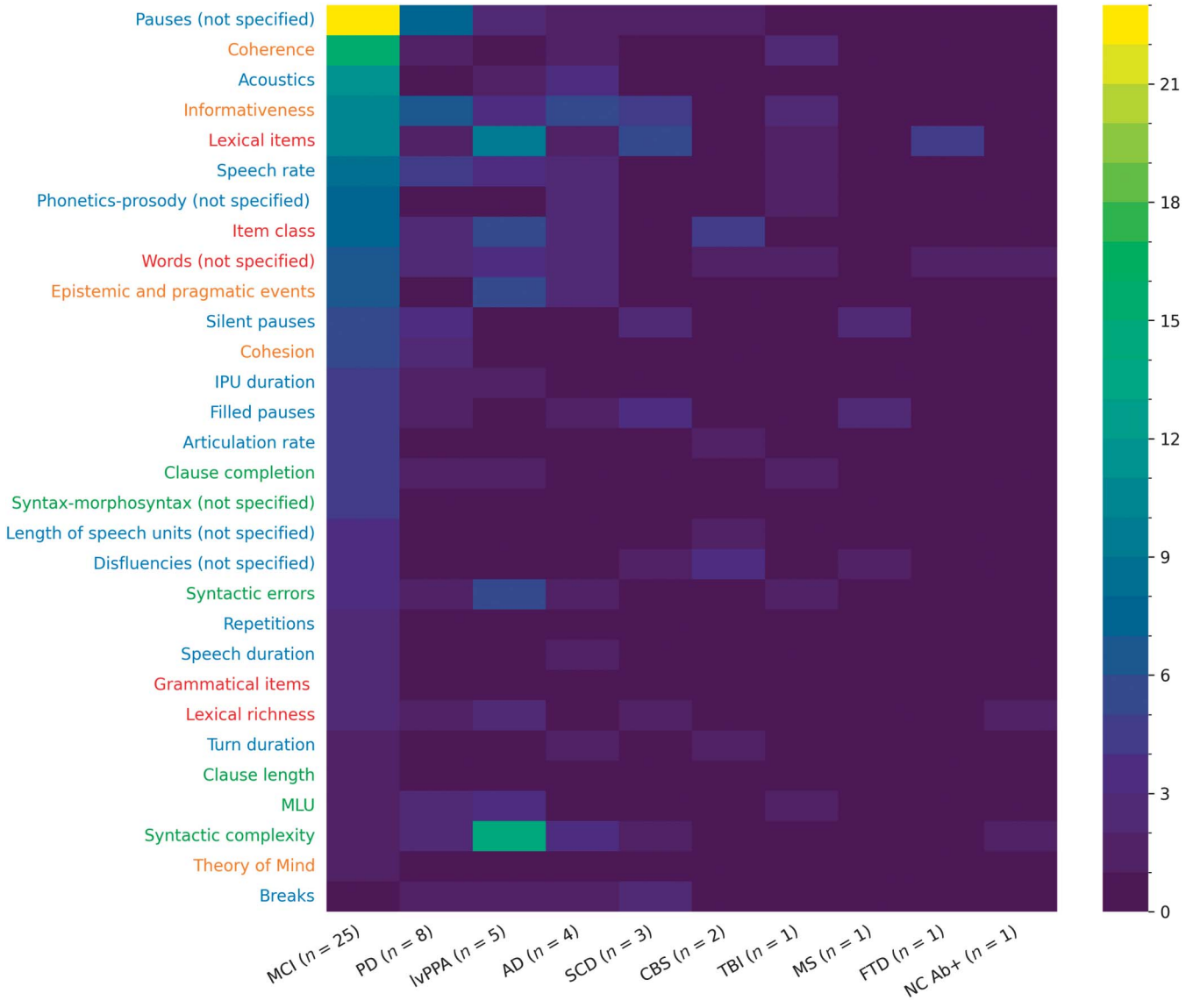
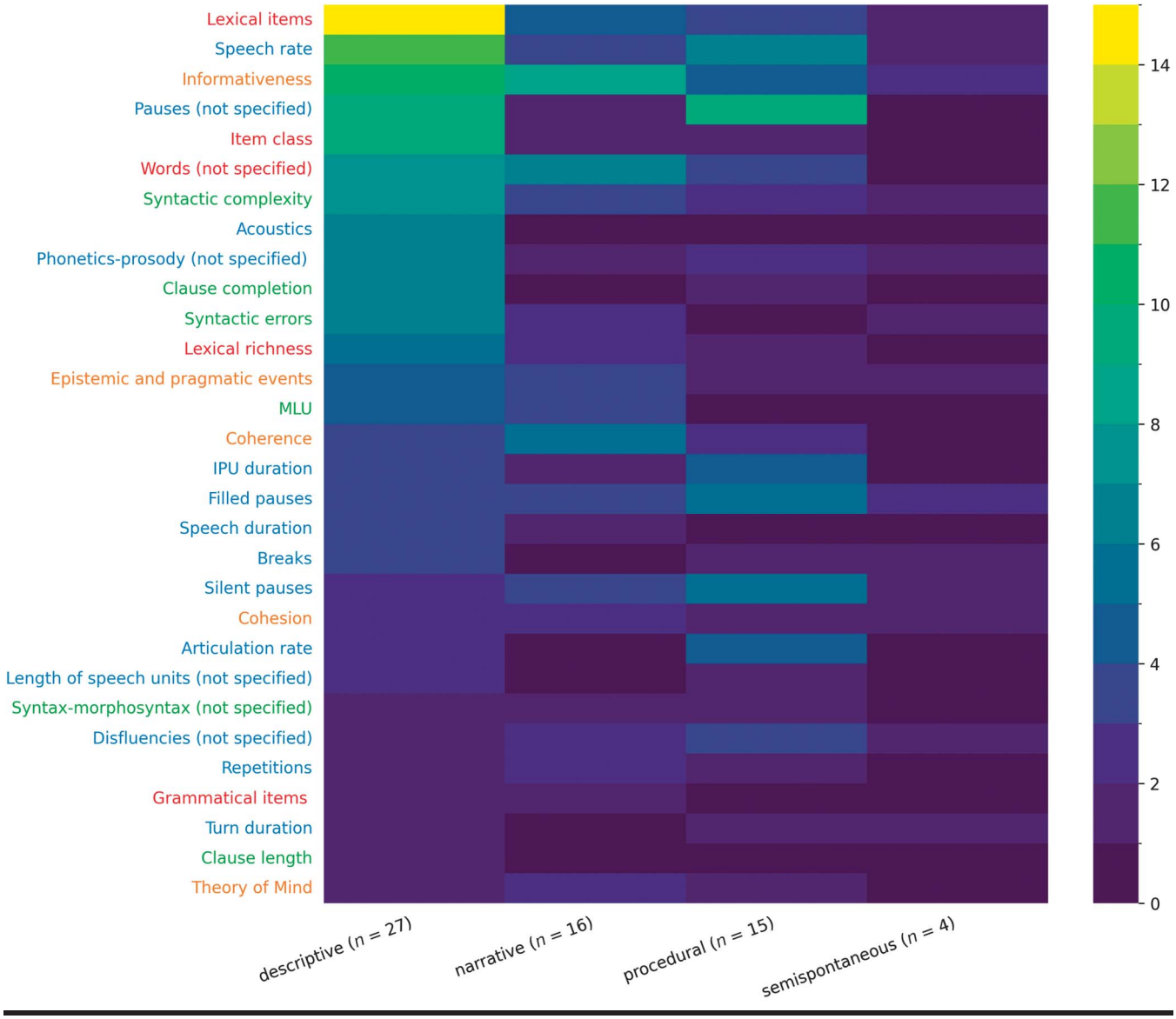


Figure 5. Linguistic feature frequency (count) according to type of discourse. Lighter colors represent higher numbers of studies. Colors of the *y*-axis labels refer to the linguistic domains: blue = phonetics-prosody; red = lexical-semantics; green = morphosyntax–syntax; orange = discourse. *n* = number of studies.



These studies mostly used pauses to investigate patients' cognition, but did not specify or provide a description of the pauses. In the same domain (i.e., phonetics-prosody), acoustic features and speech rate were also frequently used. Other commonly analyzed features belong to the domain of semantics, at the word and discourse levels (such as coherence, informativeness, and lexical items). Studies investigating logopenic variant of primary progressive aphasia (5/51) stood out in that they primarily used features related to syntax (syntactic complexity, syntactic errors) and to the lexicon (i.e., a language's inventory of lexemes with features such as lexical items, item class). However, considering the variety of features and the different ways to measure them, the range of pathologies and the small sample size, we were unable to find consistent mapping between specific features and particular phenotypes.

Figure 5 shows the frequency of use of the linguistic features depending on the type of elicited discourse. Picture description was the most frequent type of assessed discourse (27 studies), often elicited by the Cookie Theft Picture (in 18 studies). Picture description was also investigated with more features than the other discourse types. The most used linguistic features were lexical items and speech rate. The narrative and procedural discourse types were also frequently assessed (17 and 14 studies, respectively), with informativeness and pauses, respectively. Narrative discourse is

different in the use of linguistic features in that it was mainly investigated with markers belonging to word and discourse semantics but not with markers related to rhythm. Procedural discourse, on the contrary, was investigated almost exclusively with markers related to rhythm (unspecified pauses, speech rate, silent and filled pauses).

Regarding connected speech tasks, 12 studies included more than one task, and 27 studies mislabeled at least one of the connected speech tasks (i.e., featured an incorrect classification of the assessed discourse type). The descriptive and procedural types of discourse were mislabeled in 17 and 13 studies, respectively, described verbatim as either narrative or as spontaneous discourse. No study contained spontaneous speech strictly speaking (i.e., unconstrained, task-free speech in an ecological setting such as participants' homes), but four studies analyzed samples of semispontaneous discourse (i.e., speech elicited with open questions as part of semistructured interviews, which provided substantial scope for answer style or addressed topics). These four studies used filled pauses and informativeness.

Study Quality Assessment

Figure 6 shows the quality assessment ratings for each included study sorted by year of publication, while Figure 7 provides a summary of the quality assessment in function of each topic of interest.

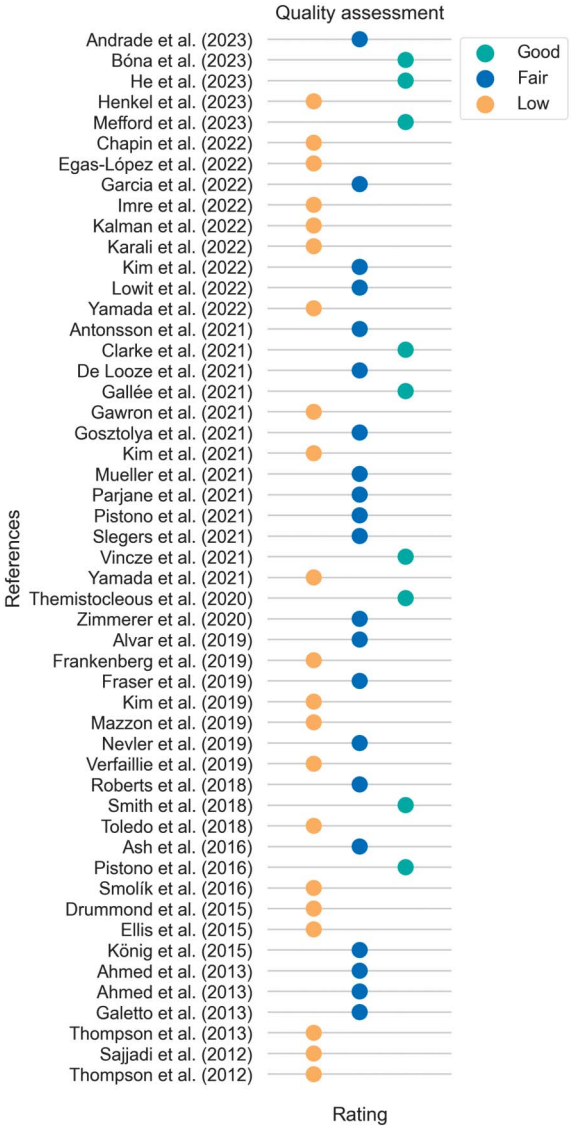
The quality assessment showed that the objectives and populations were clearly described for most of the articles. The linguistic markers and methods lacked descriptive information in more than half of the studies. The tasks involving connected speech were fully described in 37 studies out of 51. Confounding variables were not assessed in most of the studies, and 17 studies were deemed replicable despite missing information. Nine studies were rated good, 21 were rated fair, and 22 were rated low. All studies remain equally important contributions to this rapidly growing field. These ratings solely apply in the context of the scale we have designed. In no way can they be considered as an indicator of the general quality of papers.

As shown in Figure 7, the population and objectives were often clearly stated, while less than half of the studies fulfilled the criteria related to replicability. While speech tasks were detailed in three quarters of the studies, linguistic variables were fully described (i.e., the feature characteristics) in only half of them, which might impact study replicability.

Statistical Significance of Group Comparisons Via Linguistic Features

We looked at the feature *p* values that were reported in the studies to see whether some features showed consistent

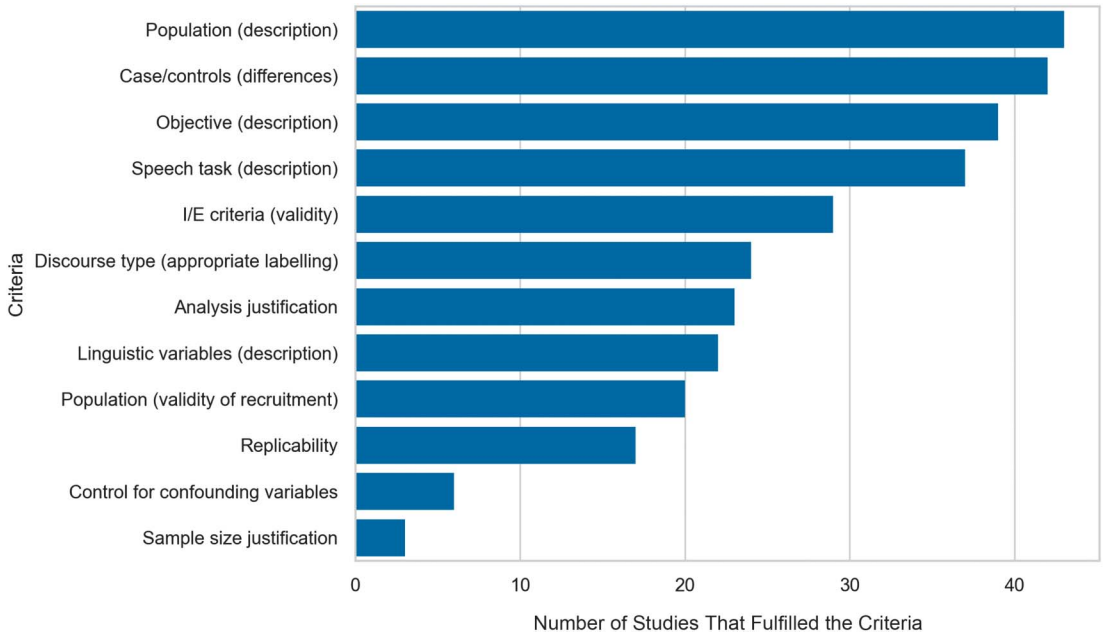
Figure 6. Quality assessment ratings.



findings independently from the underlying etiologies or behavioral-cognitive syndromes. However, the reader should keep in mind that this study gathers various etiologies together and that different results might arise from one pathology to another. Thirty-five studies used frequentist statistical tests to compare linguistic features between groups, mainly with analysis of variance (ANOVA) and Mann–Whitney *U* tests (Mann & Whitney, 1947). Across these studies, a significant *p* value or equivalent threshold regarding the statistical tests performed to compare groups was reached for 144 linguistic features out of 258 features in total. Significance probability values were not reported for six of them.

Looking at linguistic features that were used in more than 10 studies, lexical items, informativeness, and item

Figure 7. Summary of the quality assessment.



class showed a higher number of uses with significant *p* value results than uses with nonsignificant *p* values. Speech rate was used in 20 studies, with a similar number of significant and nonsignificant *p* values. Equivalent numbers of uses with significant and nonsignificant *p* values were also found for syntactic complexity and pauses (type unspecified). However, the picture is different for specified pauses, which showed very contrasting results. Silent pauses showed a lot more uses with significant *p* values, while filled pauses showed a lot fewer uses with significant *p* values. Eventually, words, that is, unspecified lexical or grammatical units, showed fewer uses with significant *p* values than uses with nonsignificant *p* values.

Summary

In terms of population, a majority of studies featured a small sample size, that is, between 10 and 29 participants. Neuropsychological assessments of the population frequently targeted executive functions as well as language and memory, but very rarely focused on attention. Picture description was the most frequent type of elicited discourse. The linguistic features we retrieved were given study-specific labels, meaning that they were very diverse in terms of names and measures (which were often unspecified). Our quality assessment also showed that more than half of the studies lacked descriptive information about the way linguistic markers were annotated and measured. The most frequently used features across studies belonged to two different domains, namely, phonetics and prosody (speech rate, pauses) and semantics (informativeness, lexical items). Pauses were also frequently analyzed, but studies did not often specify or provide a description of pause type. Results showed variability as well, with numerous features yielding equivalent numbers of significant and nonsignificant results. Lexical items, informativeness, and item class showed a higher number of reported uses with significant *p* value results than uses with nonsignificant *p* values. Speech rate was used in 20 studies, with an equivalent number of significant and nonsignificant *p* values. Equivalent numbers of uses with significant and nonsignificant *p* values were also found for syntactic complexity and pauses (type unspecified). When pause type was specified, a sharp distinction appeared between silent pauses (more uses with significant *p* values) and filled pauses (less uses with significant *p* values).

Discussion

The studies we reviewed covered a large range of populations in terms of age, etiologies, and cognitive-behavioral syndromes. The main finding of this review is the high variability in results across studies. In the first section, we relate this variability to the use of different labels for similar features and to the use of many different frequentist statistical tests for similar variable groupings and data sets. We also address the diversity in the ML models in use. In the second section, building up on the contribution of the reviewed studies to the field, we break down two different ways of assessing a feature's reliability.

The third section discusses the benefits of selecting a discourse type to elicit the function of its linguistic characteristics. Finally, the last section addresses the general implications of investigating underlying cognitive mechanisms in speech. The overall objective of this section is to discuss potential keys to reach methodological consensus while sustaining plural outlooks on subtle cognitive impairment in connected speech.

High Variability in Results

Different Labels for Similar Features

Our findings show a high diversity across studies in the linguistic features reported as reliable markers for detecting subtle cognitive impairment. This mainly stems from disparities in the way studies measured the features. Specifically, markers that were measured in different ways were reported with the same label across studies. For instance, one study may measure the feature "word number" excluding fillers and word fragments, while another study may refer to the same feature while including fillers and word fragments. This is particularly true for features that were often conflated with their measures. For instance, speech rate was one of the most frequently used markers; it is a feature that also is a measure per se. While most of the studies described speech rate as the number of words per minute, other studies reported measuring speech rate as the number of syllables per second. In addition, the use of umbrella terms yielding multiple interpretations such as "word," "utterance," or "sentence" might contribute to disparities in measurement as well. For instance, "uhm" as a marker may be classified as a word or as a pause depending on studies. Similarly, we rarely found a statement on whether word fragments (e.g., "her daugh- her kids") were counted as words, repetitions, or were removed from the count. Our findings regarding the diversity in marker measurements echo those of Boschi et al. (2017) and back their observation that a large diversity in markers hinders cross-study comparisons.

Multiplicity of Methods

Another potential cause for the variability in results lies in the multiplicity of quantitative methods used to discriminate cases from healthy controls or non-mNCD controls. In this section, we specifically address the variation found in the frequentist statistical methods and in the ML models that were used.

A large majority of studies used frequentist statistical methods. Such methods are, however, hard to implement because of the nonparametric nature of the data as well as the small sample sizes. The characteristics of language data, especially when acquired for clinical purposes, will usually entail many confounding variables such as participant age, participant treatments, or time gap between disease onset and speech recording. The suitability of classical statistical tests (i.e., ANOVAs, *t* tests) is increasingly questioned for such data (Gries, 2015). Research in statistics for health sciences suggests that mixed-effects models might provide more accurate results by controlling the confounding variables as random effects (e.g., participants; Winter & Grice, 2021). Although mixed-effects models are not as straightforward to run and interpret as other statistical methods, open-source software such as JASP now provide user-friendly linear mixed models tools (JASP Team, 2023).

Another way to discriminate mNCDs from controls is the use of ML models, which rely on the automatic classification of populations. The main advantage of using such methods relies on the number of features (sometimes more than a hundred) that can be implemented and analyzed at once, increasing the likelihood of detecting subtle cognitive change. However, the 16 studies using ML showed inconsistent model accuracies, sometimes even within a single study. While model type (e.g., SVM, random forest) may differ across studies, the heterogeneity of classifiers might not fully explain the variability in accuracy.

One other reason for the high variability in results might be related to data quality. This covers issues related to the quality of the speech output as conditioned by a specific type of elicited discourse, and issues concerning missing information or metadata. The first point has for instance been highlighted in the work of Clarke et al. (2021), who noticed that discourse type impacts model accuracy. The SVM models they ran were indeed found to perform better with features from an overlearned narrative than with features from a picture book narrative. Likewise, He et al. (2023) showed that random forest models provide better accuracies using acoustic features than transcribed features from audio to letters (sometimes called speech-to-text features). The two previous systematic reviews on classifying AD patients' speech compared to healthy controls support the idea of a better accuracy for acoustics features (de La Fuente Garcia et al., 2020; Martínez-Nicolás et al., 2021). Despite the diversity of ML models, it is important to highlight their promising use on speech assessment in clinical context.

Another significant concern relates to the lack of information on the population from which the speech sample is drawn. Clinical data are hard to access and timely to acquire. To overcome this issue, some studies used linguistic data from readily available data sets. While these data sets have the benefit of being open access (such as the Pitt Corpus; Becker et al., 1994), their main drawback is that the researcher may lack information about

the cognitive status of participants or be obliged to work with poor-quality recordings. These data flaws might lead to biases, confounding variables, and classification errors.

The impact of data quality and model types on the results assessed in this review suggests that these elements deserve more careful consideration in future studies. Researchers aiming to use ML models for clinical purposes may refer to the guidelines outlined in the work of Lo Vercio et al. (2020), where more information on the use of ML models can be found. Another useful resource is the Weka software (Frank et al., 2016) to easily run ML models, as a collection of open-source material developed by the University of Waikato, New-Zealand.

Assessing a Feature's Reliability: How Can a Feature Become a Marker?

There is no consensus regarding the metrics necessary for estimating a marker's reliability. In this section, we build up on the studies taken into account in this review and suggest considering a combination of two factors, namely, discrimination power and replicability.

Discrimination Power

The first essential factor to assess a marker's reliability is discrimination power. We define discrimination power as the number of times a feature significantly discriminates cases from controls in comparison with the number of times the feature yields nonsignificant results. In other words, discrimination power corresponds to the number of times the feature shows consistent (i.e., similar) results across studies.

This systematic review reveals that few features yielded significant results more times than they provided nonsignificant results. The high discrimination power for coherence relates to the fact that most of the studies using coherence as a marker followed consistent instructions on the evaluation of discourse coherence (see, for instance, Wright et al., 2014). Similarly, equivalent numbers of uses with significant and nonsignificant p values were found for pauses whose type was unspecified. Those pauses whose type was specified showed very contrasting results, with silent pauses showing high discrimination power, and no power for filled pauses. These findings highlight the need for clear marker definitions in order to enhance the replicability of results.

One limit of using discrimination power to assess a marker's reliability is linked to the general characteristics of language, where markers are used in contiguity with others, meaning that they interrelated. We found that subtle cognitive change in language is often interpreted through the prism of several markers analyzed in isolation. The odds for single markers to be sufficient to discriminate cases from controls are however not high. Compared to single markers in isolation, clusters of markers are much more likely to be informative in the detection of subtle cognitive impairment. For this reason, ML methods seem particularly relevant to explore cognition through linguistic markers and their patterns.

In addition to discrimination power, a marker's frequency of use is a useful indicator of its reliability. Frequency of use corresponds to the number of times the feature was used across all studies, independently from its outcome.

The results of this review showed that a majority of markers were used once, or in a small number of studies. For those features used in more than 10 studies, most were able to distinguish cases from controls approximately as often as they were unable to make this distinction. For instance, lexical items and informativeness are recurrent features that showed statistically significant outcomes in more than half of the studies using them, while speech rate is a recurrent feature that showed statistically significant outcomes in about half of the reviewed studies. Coherence, on the contrary, showed high discrimination power but was used in only eight studies out of 51. This suggests that the significance of a given marker is best assessed when weighted by its frequency of use. A feature that shows either significant or nonsignificant consistent p values across several studies is a good indicator of this feature's reliability. We suggest that features with inconsistent results should be further investigated to confirm or reject their reliability.

Replicability

One of the aims of this review was to improve the feasibility of speech analysis in clinical contexts. We found that very few studies reported having carried out an interrater agreement test as part of the cross-validation of annotations. This directly relates to the identifiability of markers, and such tests are crucial first steps for replicability (Hallgren, 2012; Holle & Rein, 2013). These findings highlight the need for harmonized methods, specifically annotation processes, when investigating speech for identifying subtle cognitive impairment.

While language may seem an accessible and relevant environment to study mNCDs, its analysis requires strong knowledge and expertise. Two major steps in speech and discourse analysis are the transcription and annotation of audio samples. These two steps are frequently done manually, which leads to many human-related biases such as auditive fatigability or false interpretation of segments. To reduce these factors, the use of semi- or fully automatized tools might be worthwhile. These systems are, however,

not perfect and need to be checked by humans. In both cases, an interrater agreement should be run to ensure the replicability of speech annotation. Replicability is ensured when the annotations performed by a researcher are consistent with those of another researcher using the same methods, without having access to the former annotations.

Linguistic Characteristics of Discourse Types

This section addresses associations between patterns of features and discourse types. Discourse types differ in their structures and contents, as well as in how their ecological validity and in the cognitive load they put on the participant. When selecting discourse type, it is therefore important to keep in mind which language skills are being assessed and, by extension, which features are more appropriate. However, the results of this systematic review must be interpreted carefully as most of the included studies focus on English speakers. Indeed, each language has its own system and rules (e.g., accentuation), and linguistic markers may vary from a language family to another one. Much research is thus needed on languages other than English to consider the difference between linguistic systems.

Marker Layout in Function of Discourse Type

The results of our clustering of the linguistic markers show that markers related to the acoustic-temporal aspects of speech (i.e., speech rate) and the lexical-semantic aspects of language (i.e., lexical items) were the most frequent across all discourse types. They also show that each discourse type was explored via a different set of markers. For instance, procedural discourse was investigated with markers focusing on the temporal aspects of discourse (e.g., pauses, length of speech units), while narratives were studied with markers related to discourse structure and semantics (e.g., informativeness, words, coherence). This finding suggests that authors attempted to select features as a function of discourse type. This approach seems particularly relevant since, in procedural discourse, participants recall elements or learned tasks in a sequential order, while in narrative tasks, participants build a story incorporating elements of their own interpretation, using markers of coherence and cohesion.

This contrast does not hold for descriptive discourse, for which markers were more varied and belonged to all linguistic domains. The wider range of markers may be explained by the overrepresentation of descriptive speech. Description tasks are indeed easier to control and standardize. However, they are far from ecological, as describing entities on pictures does not reflect a participant's everyday language use in context. As highlighted by Bryant et al. (2016), this raises the question of whether descriptive discourse is a good candidate for revealing/investigating subtle cognitive impairment as encountered by patients in their daily lives. Contrastively, spontaneous, or specifically semispontaneous, discourse feature in very few investigations. While this discourse type is the closest to everyday language, it might also be the hardest to investigate, with many confounding variables.

A related point in question is that of the lack of harmonization regarding nomenclature on discourse types. Several studies sought to assess spontaneous speech, but used a picture-based description task, meaning that their results were uninformative with respect to their aim. Such mismatches might lead to strong biases and might create a focus on markers that are irrelevant to spontaneous speech.

In addition to previous research suggesting to mix discourse types to get a good overview of language skills (Brookshire & Nicholas, 1994), we advocate selecting different markers for different discourse types. We hypothesize that targeting specific markers sheds better light on the abilities and frailties of the population with mNCDs.

Exploring Cognition Through Speech Analysis

The final point of interest in this review relates to the selection of linguistic markers to investigate the underlying cognitive mechanisms of mNCDs. We specifically discuss the fact that few of the reviewed studies explored patients' cognition in function of discourse type and in function of the markers they focused on.

Hypotheses on Underlying Mechanisms of Discourse Type

Language use is cognitively complex in that it not only involves specific processing mechanisms such as semantics and phonology but also relies on other cognitive functions such as memory, executive functions (including working memory [Baddeley, 2003] and processing speed), as well as attentional processes. The discourse type elicited by the speech production task will draw upon these cognitive functions in specific ways and intensities. For instance, narratives may require memory more than descriptive discourse. The output speech variables might also vary depending on discourse type. As suggested by He et al. (2023), morphological features are likely to be associated with procedural memory and therefore may not be relevant for people with short-term or semantic memory deficits such as individuals with AD or MCI. Research in neuroscience on language learning has shown the implication of procedural memory in the implicit learning of linguistic sequences such as morphology and syntax rules (Nemeth et al., 2011; Ullman, 2016).

This can guide the selection of a particular type of discourse over another, depending on the targeted cognitive

functions and the clinical context. Yet, we found very little justification about the choice of speech task relative to underlying cognitive functions, which suggests that such relations between discourse type and cognitive functions represent a promising avenue for future studies.

Neuropsychological and Neural Correlates of Linguistic Markers

When asked about their cognition, patients with mNCDs mostly report patterns of difficulties that cover a variety of cognitive domains (Verfaillie et al., 2019). It is therefore not surprising that executive functions, memory, and/or language are often assessed altogether by authors. Some of the reviewed studies tested the presence of a correlation between speech markers and neuropsychological scores, and found significant results (see, e.g., Parjane et al., 2021; Smith et al., 2018; Zimmerer et al., 2020). Such findings may help generate hypotheses on underlying cognitive impairment. For instance, De Looze et al. (2021) performed linear mixed-effects models to examine whether the test scores of an isolated cognitive function (e.g., working memory) could predict temporal features in speech (e.g., number of pauses). In the procedural discourse of MCI patients, they observed associations between speech rate and memory, turn duration and working memory/attention, and interpausal units (IPUs) and memory. They suggest that a slower speech rate and a longer duration for pauses might signal deficits in episodic memory as well as lexical, semantic, and executive functioning processes.

Yet, these studies selected different markers along with different tests, and ran mixed models or linear regressions. Replicating such promising results that map validated test scores onto linguistic markers is thus a long journey given the variability in methods. Furthermore, the small number of studies that have investigated associations between speech and neuropsychological variables means that the interpretation of such findings will benefit from further research.

In addition to neuropsychological assessments, some studies used neuro-imagery to investigate potential associations between linguistic features and brain regions. Pistono et al. (2021) used both structural and functional magnetic resonance imaging (fMRI) in MCI patients versus healthy controls during a picture-based narrative task to explore regions related to language and to executive functions. Although they found significant differences between groups for speech markers, they found no correlation between language network structure or functionality and language performance despite reporting a significant decrease in language gray matter compared to controls. De Looze et al. (2021) investigated the structural correlates of procedural features in MCI, targeting nine regions of interest (ROIs) involved in speech production. Combining both neuropsychological assessment and fMRI techniques, they found interactions between ROIs and speech rate, turn duration, and IPU duration. The authors suggest that in addition to revealing lexical and semantic deficits, temporal features may reflect planning difficulties in speech production due to damaged working memory and attention skills.

Limitations of This Review and Future Research

This systematic review targeted a wide array of pathologies and cognitive-behavioral syndromes to better reflect an ongoing clinical reality. The results of this review are therefore relevant for an overview of subtle cognitive impairment. The literature on linguistic markers of mNCDs is likely to increase the next few years. We suggest that future research should focus on subsets of pathologies or cognitive-behavioral syndromes. In particular, there is a need to focus on underrepresented pathologies, such as multiple sclerosis, to better grasp the relationship between the linguistic features and the underlying cognitive mechanisms.

One bias of this review is linked to the fact that we analyzed linguistic markers based on our labeling system. While this method offers a wide linguistic perspective, it might prevent a perfect match between the study results and ours. For instance, "informativeness" as a label encompasses features such as "idea density" and "information content," while a distinction between them was made by some authors. However, we estimate that our method does not impact the readability of our results since we provided a description of the labels we used.

This study has included both manual and automatic analyses of speech. Another limitation concerns automatic analyses, in that we did not rely on ML-specific criteria to further compare these methods and models. Further studies covering specific aspects in ML methods for detecting subtle cognitive impairment are thus necessary.

Since the majority of the reviewed studies focused on English, we were unable to draw any conclusion regarding possible differences in linguistic markers between language. Another study with a special focus on cross-languages comparison is thus needed.

This review was designed to offer a number of starting points for further research in clinical and language science to gain better insight on linguistic markers of subtle cognitive impairment. However, the detection of mNCDs through speech is a rapidly evolving field, as new articles have been published since our last database search. We provide our working definitions for features, data tables,

and assessment grids as their value and relevance can persist through their reuse.

Contributions of This Review

This review covers different fields of interest. In this section, we state what this review adds to previous literature depending on the reader's background and interests.

1. Interest in analyzing connected speech:
   - A list of currently examined linguistic markers is available in the Results section, as well as our clustering approach for categorizing linguistic features in the Method section.
   - A detailed overview of the features and their definitions are proposed in Table 2.
   - The Exploring Cognition Through Speech Analysis section in the Discussion section presents an opinion on current challenges in the domain of speech analysis.
2. Interest in assessing cognition using a discourse task:
   - The Method section draws up a list of the various pathologies and cognitive-behavioral syndromes we included.
   - The variability in feature outcomes is revealed in the Results section.
   - The Discussion section provides an overview of the association between linguistic markers and cognitive impairment, as this will provide guidance when choosing features and discourse tasks.

Conclusions

Our systematic review targeted mNCDs, which covers a large range of underlying pathologies characterized by a subtle decline in cognitive performance. This review aimed at surveying the speech features in use to identify those that can best detect subtle cognitive impairment in individuals with mNCDs, whose prevalence is increasing. The underlying physiopathology of such subtle cognitive changes often remain undiagnosed. Patients complain of cognitive impairment, which can hardly be detected by current neuropsychological tools. In order to improve the reliability of markers, future studies should specify the markers they used, as well as the type of discourse they elicited along with the methods to annotate and analyze the data. This systematic review highlights the need for a harmonized terminology. A first step toward this is a detailed description of linguistic markers to improve the interpretation and the generalization of studies. A further development of this review is therefore to provide a detailed paper on harmonized labels and definitions of the linguistic markers in connected speech.

Author Contributions

Amélie B. Richard: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Manon Lelandais:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Sophie Jacquin-Courtois:** Conceptualization, Writing – review & editing, Supervision. **Karen T. Reilly:** Conceptualization, Writing – review & editing, Supervision.

Data Availability Statement

The appendices and our classification system are available on Open Science Framework (OSF): https://osf.io/qhtzp/?view_only=249a568932d24cf7986159df2a0eb30d.

Acknowledgments

Amélie B. Richard is financed by a PhD grant awarded by Paul-Valery Montpellier 3 University and funded by the French Ministry of Higher Education (Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation).

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (DSM-5-TR)* (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Andrade, E. I. N., Manxhari, C., & Smith, K. M. (2023). Pausing before verb production is associated with mild cognitive impairment in Parkinson's disease. *Frontiers in Human Neuroscience, 17,* Article 1102024. https://doi.org/10.3389/fnhum.2023.1102024

Areklett, E. W., Andersson, S., Fagereng, E., Bruheim, K., Stubberud, J., & Lindemann, K. (2024). Cognitive impairment in cervical cancer survivors—Exploring the discrepancy between subjective and objective assessment. *Psycho-Oncology, 33*(2), Article e6300. https://doi.org/10.1002/pon.6300

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

Belz, M. (2023). Defining filler particles: A phonetic account of the terminology, form, and grammatical classification of "filled pauses." *Language, 8*(1), Article 57. https://doi.org/10.3390/languages8010057

Bigi, B., & Priego-Valverde, B. (2022). The automatic search for sounding segments of SPPAS: Application to Cheese! Corpus. In Z. Vetulani, P. Paroubek, & M. Kubis (Eds.), *Human language technology. Challenges for computer science and linguistics* (Vol. 13212, pp. 16–27). Springer International Publishing. https://doi.org/10.1007/978-3-031-05328-3_2

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.1.55) [Computer software]. https://www.praat.org

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,* Article 269. https://doi.org/10.3389/fpsyg.2017.00269

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., & Blache, P. (2021). A multimodal model for predicting conversational feedbacks. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), *Text, speech, and dialogue* (Vol. 12848, pp. 537–549). Springer International Publishing. https://doi.org/10.1007/978-3-030-83527-9_46

Brookshire, R. H., & Nicholas, L. E. (1994). Test–retest stability of measures of connected speech in aphasia. *Clinical Aphasiology, 22,* 119–133. http://aphasiology.pitt.edu/id/eprint/163

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics, 30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Clarke, N., Barrick, T. R., & Garrard, P. (2021). 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,* Article 44. https://doi.org/10.3389/fcomp.2021.634360

Costa, D. S. J., & Fardell, J. E. (2019). Why are objective and perceived cognitive function weakly correlated in patients with cancer? *Journal of Clinical Oncology, 37*(14), 1154–1158. https://doi.org/10.1200/JCO.18.02363

Cuenca, M. J., & Crible, L. (2019). Co-occurrence of discourse markers in English: From juxtaposition to composition. *Journal of Pragmatics, 140,* 171–184. https://doi.org/10.1016/j.pragma.2018.12.001

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

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. https://doi.org/10.3233/JAD-200888

De Looze, C., Dehsarvi, A., Crosby, L., Vourdanou, A., Coen, R. F., Lawlor, B. A., & Reilly, R. B. (2021). Cognitive and structural correlates of conversational speech timing in mild cognitive impairment and mild-to-moderate Alzheimer's disease: Relevance for early detection approaches. *Frontiers in Aging Neuroscience, 13,* Article 637404. https://doi.org/10.3389/fnagi.2021.637404

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech, 25*(1), 11–28. https://doi.org/10.1177/002383098202500102

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

Fox Tree, J. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*(6), 709–738. https://doi.org/10.1006/jmla.1995.1032

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, Fourth Edition, 2016*. The University of Waikato.

Galetto, V., Andreetta, S., Zettin, M., & Marini, A. (2013). Patterns of impairment of narrative language in mild traumatic brain injury. *Journal of Neurolinguistics, 26*(6), 649–661. https://doi.org/10.1016/j.jneuroling.2013.05.004

Gries, S. Th. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics, 16*(1), 93–117. https://doi.org/10.1177/1606822X14556606

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

He, R., Chapin, K., Al-Tamimi, J., Bel, N., Marquié, M., Rosende-Roca, M., Pytel, V., Tartari, J. P., Alegret, M., Sanabria, A., Ruiz, A., Boada, M., Valero, S., & Hinzen, W. (2023). Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains. *American Journal of Speech-Language Pathology, 32*(5), 2075–2086. https://doi.org/10.1044/2023_AJSLP-22-00403

Henderson, F. M., Cross, A. J., & Baraniak, A. R. (2019). 'A new normal with chemobrain': Experiences of the impact of chemotherapy-related cognitive deficits in long-term breast cancer survivors. *Health Psychology Open, 6*(1). https://doi.org/10.1177/2055102919832234

Holle, H., & Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. In H. Lausberg (Ed.), *Understanding body movement: A guide to empirical research on nonverbal behaviour (with an introduction to the NEUROGES coding system)* (pp. 261–275). Peter Lang Verlag.

Ivanova, O., Martínez-Nicolás, I., & Meilán, J. J. G. (2023). Speech changes in old age: Methodological considerations for speech-based discrimination of healthy ageing and Alzheimer's disease. *International Journal of Language & Communication Disorders, 59*(1), 13–37. https://doi.org/10.1111/1460-6984.12888

JASP Team. (2023). *JASP* (Version 0.17.3) [Computer software].

Kehler, A. (2006). Discourse coherence. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (1st ed., pp. 241–265). Wiley. https://doi.org/10.1002/9780470756959.ch11

Labov, W. (2001). Uncovering the event structure of narrative. In D. Tannen & J. Alatis (Eds.), *Georgetown University round table on languages and linguistics 2001* (pp. 63–83). Georgetown University Press.

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M. L. (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology, 32*(2), 426–438. https://doi.org/10.1044/2022_AJSLP-22-00281

Lo Vercio, L., Amador, K., Bannister, J. J., Crites, S., Gutierrez, A., MacDonald, M. E., Moore, J., Mouches, P., Rajashekar, D., Schimert, S., Subbanna, N., Tuladhar, A., Wang, N., Wilms, M., Winder, A., & Forkert, N. D. (2020). Supervised machine learning tools: A tutorial for clinicians. *Journal of Neural Engineering, 17*(6), Article 062001. https://doi.org/10.1088/1741-2552/abbff2

Lowit, A., Thies, T., Steffen, J., Scheele, F., Roheger, M., Kalbe, E., & Barbe, M. (2022). Task-based profiles of language impairment and their relationship to cognitive dysfunction in Parkinson's disease. *PLOS ONE, 17*(10), Article e0276218. https://doi.org/10.1371/journal.pone.0276218

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the

other. *Annals of Mathematical Statistics, 18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

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,* Article 620251. https://doi.org/10.3389/fpsyg.2021.620251

Mateo, S. (2020). Procédure pour conduire avec succès une revue de littérature selon la méthode PRISMA [Using the PRISMA method to conduct a literature review]. *Kinésithérapie, la Revue, 20*(226), 29–37. https://doi.org/10.1016/j.kine.2020.05.019

McDonald, W. M. (2017). Overview of neurocognitive disorders. *Focus, 15*(1), 4–12. https://doi.org/10.1176/appi.focus.20160030

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330. https://doi.org/10.1080/01638530902959943

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41*(4), 215–225. https://doi.org/10.1159/000261728

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *PLOS Medicine, 6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nemeth, D., Janacsek, K., Csifcsak, G., Szvoboda, G., Howard, J. H., & Howard, D. V. (2011). Interference between sentence processing and probabilistic implicit sequence learning. *PLOS ONE, 6*(3), Article e17577. https://doi.org/10.1371/journal.pone.0017577

Pallaud, B., Bertrand, R., Blache, P., Prévot, L., & Rauzy, S. (2019). Suspensive and disfluent self interruptions in French language interactions. In L. Degand, G. Gilquin, L. Meurant, & A.-C. Simon (Eds.), *Fluency and disfluency across languages and language varieties* (pp. 109–138). Presses universitaires de Louvain.

Parjane, N., Cho, S., Ash, S., Cousins, K. A. Q., Shellikeri, S., Liberman, M., Shaw, L. M., Irwin, D. J., Grossman, M., & Nevler, N. (2021). Digital speech analysis in progressive supranuclear palsy and corticobasal syndromes. *Journal of Alzheimer's Disease, 82*(1), 33–45. https://doi.org/10.3233/JAD-201132

Pati, D., & Lorusso, L. N. (2018). How to write a systematic review of the literature. *Health Environments Research & Design Journal, 11*(1), 15–30. https://doi.org/10.1177/1937586717747384

Petersen, R. C. (2016). Mild cognitive impairment. *Continuum, 22*(2, Dementia), 404–418. https://doi.org/10.1212/CON.0000000000000313

Pistono, A., Senoussi, M., Guerrier, L., Rafiq, M., Giméno, M., Péran, P., Jucla, M., & Pariente, J. (2021). Language network connectivity increases in early Alzheimer's disease. *Journal of Alzheimer's Disease, 82*(1), 447–460. https://doi.org/10.3233/JAD-201584

Röhr, S., Pabst, A., Riedel-Heller, S. G., Jessen, F., Turana, Y., Handajani, Y. S., Brayne, C., Matthews, F. E., Stephan, B. C. M., Lipton, R. B., Katz, M. J., Wang, C., Guerchet, M., Preux, P.-M., Mbelesso, P., Ritchie, K., Ancelin, M.-L., Carrière, I., Guaita, A., . . . Sachdev, P. S. (2020). Estimating prevalence of subjective cognitive decline in and across international cohort studies of aging: A COSMIC study. *Alzheimer's Research & Therapy, 12*(1), 167. https://doi.org/10.1186/s13195-020-00734-y

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735. https://doi.org/10.1016/B978-0-12-623550-0.50008-2

Schiffrin, D. (1987). *Discourse markers* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511611841

Smith, K. M., Ash, S., Xie, S. X., & Grossman, M. (2018). Evaluation of linguistic markers of word-finding difficulty and cognition in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 61*(7), 1691–1699. https://doi.org/10.1044/2018_JSLHR-L-17-0304

Tangalos, E. G., Smith, G. E., Ivnik, R. J., Petersen, R. C., Kokmen, E., Kurland, L. T., Offord, K. P., & Parisi, J. E. (1996). The Mini-Mental State Examination in general medical practice: Clinical utility and acceptance. *Mayo Clinic Proceedings, 71*(9), 829–837. https://doi.org/10.4065/71.9.829

Ullman, M. T. (2016). The declarative/procedural model. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 953–968). Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Van Kuppevelt, J. (1995). Main structure and side structure in discourse. *Lingua, 33*(4), 809–833. https://doi.org/10.1515/ling.1995.33.4.809

Verfaillie, S. C. J., Witteman, J., Slot, R. E. R., Pruis, I. J., Vermaat, L. E. W., Prins, N. D., Schiller, N. O., van de Wiel, M., Scheltens, P., van Berckel, B. N. M., van der Flier, W. M., & Sikkes, S. A. M. (2019). High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline. *Neuropsychologia, 131,* 184–192. https://doi.org/10.1016/j.neuropsychologia.2019.05.006

Vos, L., Williams, M. W., Poritz, J. M. P., Ngan, E., Leon-Novelo, L., & Sherer, M. (2020). The discrepancy between cognitive complaints and neuropsychological test findings in persons with traumatic brain injury. *Journal of Head Trauma Rehabilitation, 35*(4), E382–E392. https://doi.org/10.1097/HTR.0000000000000557

Winter, B., & Grice, M. (2021). Independence and generalizability in linguistics. *Linguistics, 59*(5), 1251–1277. https://doi.org/10.1515/ling-2019-0049

Wright, H. H., Koutsoftas, A. D., Capilouto, G. J., & Fergadiotis, G. (2014). Global coherence in younger and older adults: Influence of cognitive processes and discourse type. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 21*(2), 174–196. https://doi.org/10.1080/13825585.2013.794894

Zimmerer, V. C., Hardy, C. J. D., Eastman, J., Dutta, S., Varnet, L., Bond, R. L., Russell, L., Rohrer, J. D., Warren, J. D., & Varley, R. A. (2020). Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: An approach based on usage-frequency. *Cortex, 133,* 103–119. https://doi.org/10.1016/j.cortex.2020.08.027