ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Automatically Calculated Context-Sensitive Features of Connected Speech Improve Prediction of Impairment in Alzheimer's Disease

Graham Flick[a,b] and **Rachel Ostrand**[c]

[a]Department of Psychology, New York University, NY [b]Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada [c]IBM Research, Yorktown Heights, NY

ARTICLE INFO

Article History:
Received May 3, 2024
Revision received December 2, 2024
Accepted July 11, 2025

Editor-in-Chief: Julie A. Washington
Editor: Jacqueline J. Hinckley

https://doi.org/10.1044/2025_JSLHR-24-00297

ABSTRACT

Purpose: Early detection is critical for effective management of Alzheimer's disease (AD) and other dementias. One promising approach for predicting AD status is to automatically calculate linguistic features from open-ended connected speech. Past work has focused on individual word-level features such as part of speech counts, total word production, and lexical richness, with less emphasis on measuring the relationship between words and the context in which they are produced. Here, we assessed whether linguistic features that take into account where a word was produced in the discourse context improved the ability to predict AD patients' Mini-Mental State Examination (MMSE) scores and classify AD patients from healthy control participants.
Method: Seventeen linguistic features were automatically computed from transcriptions of spoken picture descriptions from individuals with probable or possible AD (n = 176 transcripts). This included 12 word-level features (e.g., part of speech counts) and five features capturing contextual word choices (linguistic surprisal, computed from a computational large language model, and properties of words produced following filled pauses). We examined whether (a) the full set jointly predicted MMSE scores, (b) the addition of contextual features improved prediction, and (c) linguistic features could classify AD patients (n = 130) versus healthy participants (n = 93).
Results: Linguistic features accurately predicted MMSE scores in individuals with probable or possible AD and successfully identified up to 87% of AD participants versus healthy controls. Statistical models that contained linguistic surprisal (a contextual feature) performed better than those that included only word-level and demographic features. Overall, AD patients with lower MMSE scores produced more empty words, fewer nouns and definite articles, and words that were higher frequency yet more surprising given the previous context.
Conclusion: These results provide novel evidence that metrics related to contextualized word choices, particularly the surprisal of an individual's words, capture variance in degree of cognitive decline in AD.

It is estimated that by 2050, over 115 million people worldwide will be living with dementia, a group of cognitive symptoms that includes declines in memory, language, and problem-solving, as well as difficulty concentrating, confusion, and struggling to express one's thoughts (Alzheimer's Association, 2023; Prince et al., 2013). Alzheimer's disease (AD), the progressive neurological disease that is the most common cause of dementia, is itself expected to double in prevalence by 2060 (Alzheimer's Association, 2023). As there is currently no cure, the impacts of dementia, both from AD and from other sources, may be better managed through earlier and more accurate diagnoses (Black et al., 2017; Rasmussen & Langerman, 2019; Siemers et al., 2016). Early detection,

Correspondence to Graham Flick: graham.flick@nyu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

especially the distinction between dementia and cognitive changes typical in aging, may also significantly reduce the burden on caregivers, providing more time to adjust to the patient's potential changes in mood, personality, and dependence that are associated with AD pathology (de Vugt & Verhey, 2013; Mittelman et al., 1996). Unfortunately, however, existing screening methods are often costly and require substantial expertise to administer; they may be insensitive to mild changes in cognitive abilities that could indicate early stages of AD; and they are frequently only able to be administered in clinical settings, which can require prohibitive amounts of time and/or travel for some individuals to access (Bayly et al. 2020; Bradford et al., 2009; Wilson et al., 2023).

Recently, there has been increased focus on developing complementary screening tools that could address these shortcomings. One approach that has shown promise in many studies is the use of linguistic features computed from open-ended connect speech (e.g., Bucks et al., 2000; Fraser et al., 2015; Guinn & Habash, 2012; Jarrold et al., 2014; Meilán et al., 2014; Thomas et al., 2005; see Mueller, Hermann, et al., 2018, for a review). Foundational studies demonstrated that cognitive decline in AD is associated with difficulties in word-finding and semantic retrieval, related to the degradation of meanings in memory and a poorer ability to access them (see Burke & Shafto, 2008; Kemper & Altmann, 2017). This is posited to explain attested patterns in mild-to-moderate AD patients' connected language, which include reduced semantic specificity (Forbes-McKay et al., 2013; Nicholas et al., 1985) and greater use of anaphora, indefinite articles, and high-frequency words (Croisile et al., 1996; Hier et al., 1985; Kavé & Dassa, 2018; Nicholas et al., 1985; Slegers et al., 2018).

Even in early stages of dementia, such as among asymptomatic individuals with genetic AD risk factors (Cuetos et al., 2007) or early mild cognitive impairment due to AD (Ahmed et al., 2013; Mueller, Koskic, Hermann, et al., 2018; Mueller et al., 2016), connected speech has been suggested to show signs of the disease's impact. That is, relative to healthy controls (see Mueller, Koscik, Clark, et al., 2018), these individuals' speech contains less semantic content—including a reduction in relevant vocabulary items in picture descriptions (Stark et al., 2025)—and reductions in fluency (i.e., more repetitions, revisions, and filled or unfilled pauses; Mueller, Koscik, Hermann, et al., 2018). These changes, particularly in semantic content, appear to become more pronounced with disease progression (Ahmed et al., 2013). Other studies have found that AD progression also co-occurs with more frequent production of words and phrases that are not expected in the current context (Croisile et al., 1996; Forbes et al., 2002; Murray, 2010) and the need for more speaking turns to convey an intended message (Feyereisen et al., 2007).

As others have noted (e.g., Fraser et al., 2015; Mueller, Hermann, et al., 2018), these changes in connected speech patterns suggest promise in identifying and tracking stages of cognitive decline, particularly in prodromal AD where they could distinguish between changes due to typical aging. One reason is that these speech and language effects often occur prior to more overt memory deficits and the degradation of more holistic language processing, which causes obvious impacts on coherent communication (such as that caused by deficits in articulatory production and syntax; Bayles et al., 1992; Filiou et al., 2019; Forbes-McKay et al., 2013; Hier et al., 1985; Mueller et al., 2016). Moreover, connected or spontaneous speech samples can be collected on a regular basis with little burden on the individual participant through, for example, telephone interviews (Diaz-Asper et al., 2021). These samples and calculated linguistic features could thus be used as an early warning system to trigger a thorough clinical assessment, as individual-specific baselines to measure the trajectory of disease progression, and/or to provide informed strategies for dealing with communication loss—one of the most frequently reported difficulties that caregivers face (Murray et al., 1999). Indeed, prior work has demonstrated that automatically extracted linguistic features—collected during picture descriptions and monologues—explain significant variability in older adults' performances on neuropsychological assessments, including those conducted 1 year into the future (Ostrand & Gunstad, 2021).

The present investigation builds on this prior work by testing the utility of two new classes of linguistic features in predicting Mini-Mental State Examination (MMSE) scores and classifying AD patients from healthy controls. Several past studies using fully automated methods have examined the relationship between the presence or severity of AD and linguistic features derived from (a) properties and counts of individual words or word types, such as overall lexical frequency and the count of various parts-of-speech (e.g., nouns, verbs); (b) measures of speech fluency (e.g., counts of repetitions, revisions, filler words); and (c) measures of lexical richness, such as the type–token ratio (ratio of unique words to total words; Blanken et al., 1987; Bucks et al., 2000) and Honoré's statistic (a measure of words produced only once; Honoré, 1979). Each of these metrics is based on the occurrence or properties of individual words, without regard to the word's place within the larger discourse context. In other words, past work has focused on linguistic features that have the same value regardless of individual words' position in the discourse: *Dog* is counted as a noun[1] and as a specific lexical frequency, and the transcript has the same ratio of types to

[1]The exception to this is if a word's position results in a change to its syntactic class or assigned part of speech (e.g., *dog* used as a verb).

tokens, regardless of whether words occur in predicted locations ("I have a pet *dog*"), the transcript is scrambled ("pet a have I *dog*"), or the participant produces the word in an unpredictable context ("I mailed a coconut to my *dog*"). In contrast, relatively little attention has been placed on investigating word production in relation to the context where the word was produced in the discourse—features for which the position of an individual word does matter—and whether features that capture this information are meaningful for differentiating patients with dementia from healthy controls or predicting cognitive test scores.

Here, we considered two classes of these contextually informed linguistic features, referred to as "context features" for short, which we hypothesized might change in AD and with the disease's progression. The first is the properties of words that directly followed filler words (e.g., *um, ah*), as fillers are often considered a signal of difficulty in lexical retrieval or deciding how to continue the sentence (Clark & Fox Tree, 2002). We hypothesized AD patients would produce filler words in different contexts than healthy controls and that this behavior would increase with disease progression. Specifically, we expected that AD patients would have greater difficulty with lexical retrieval and thus display (a) higher median frequency of words that followed a filler and (b) a greater distance between that filler and the next content-carrying word.

Second, we took advantage of computational language modeling to estimate the probability of individual words given the context they appeared in. This enabled us to develop measures based on linguistic surprisal (Shannon, 1948) of how unexpected a speaker's words were within the ongoing discourse. Based on the past findings that AD progression is related to declines in speech fluency (i.e., increases in revisions and repetitions; Mueller, Koscik, Hermann, et al., 2018) and increases in empty words (Hier et al., 1985; Nicholas et al., 1985) or words that are contextually unexpected or inappropriate (Croisile et al., 1996; Forbes et al., 2002; Murray, 2010), we hypothesized that AD patients would show overall increases in the linguistic surprisal of their words and that this would heighten with disease progression. We estimated the word-by-word surprisal of each speaker's speech sample using the pretrained GPT-2 computational language model (Radford et al., 2019). This model captures the statistical tendencies of the English language through its training in next-word prediction from over 8 million online documents. As such, unexpected word choices or departures from fluency would result in higher overall estimated surprisal values, providing a means to operationalize these word choice patterns in a single variable.

To test the relevance of these novel properties as diagnostic signals for AD and compare their efficiency to a larger set of linguistic features from prior literature, we analyzed picture descriptions and MMSE scores collected as part of the Pitt corpus of DementiaBank (Becker et al., 1994). This resource includes data from patients with probable and possible AD, as well as healthy controls. While MMSE scores are not a perfect measure of one's degree of cognitive decline due to AD (Arevalo-Rodriguez et al., 2015, 2021), we considered this to be a first step in evaluating the new features, which, if successful, would motivate follow-up work. All linguistic features were calculated from transcriptions of picture descriptions using an automated pipeline, thus requiring no manual coding. Once computed, we examined the relationship between the linguistic features and dementia in AD in two ways. First, we investigated whether the computed features could predict the AD patients' MMSE scores and whether the novel, contextually informed features we developed here explained additional variance, beyond that captured by previously studied properties. In this way, we tested whether the new features improved prediction of an available proxy measure of dementia severity (i.e., cognitive impairment) in AD patients. Second, we investigated whether the overall set of linguistic features could be used for binary classification by building a model to separate participants into healthy controls versus AD patients, based solely on features computed from their connected speech production.

Method

Data Set

The data analyzed in the current study come from the Pitt corpus of DementiaBank, collected as part of the Alzheimer Research Program at the University of Pittsburgh (Becker et al., 1994) and made publicly available through the TalkBank corpus (http://www.talkbank.org). The current study was not subject to an approval process as it consisted of a reanalysis of existing data and no new data collection. The original study sample consisted of 204 AD patients and 102 healthy control participants, recruited from a variety of sources, including the Allegheny County Medical society, Pittsburgh-area neurologists and psychiatrists, and direct patient referral from the Benedum Geriatric Center at the University of Pittsburgh Medical Center. To be included in the original study, participants were required to be at least 44 years of age with the ability to read and write in English (prior to dementia onset), have completed education beyond the seventh grade, have no history of nervous system disorders or having regularly taken medications affecting the central nervous system (excluding antidepressants), and have successfully completed the MMSE (Folstein, Folstein, & McHugh, 1975) and

achieved an initial score of 10 or greater (test range: 0–30). Once enrolled, all participants underwent an extensive neuropsychiatric evaluation over the course of 2 weeks, which included a detailed medical history and physical exam, neurologic history and examination, semistructured psychiatric interview, and a neuropsychological assessment. Participants also completed laboratory tests, including a standard battery of hematologic studies, blood chemistry studies, liver and thyroid function tests, vitamin level assessments, a rapid plasma regain test, an electroencephalogram, and a computed tomography scan of the head.

The results of these evaluations were reviewed by the original study's team of clinicians to classify participants as AD patients or healthy controls. Criteria for classification as AD included that the individual demonstrate "a history of progressive cognitive and functional decline and an abnormal mental status examination (performed during the neurological examination)." After the study closed, the clinical records of individual participants, as well as autopsies if available, were reviewed to arrive at a final, consensus clinical classification.[2] This resulted in a final sample of 181 individuals classified by the clinical team as exhibiting probable AD. Neuropsychological test data collected from these individuals confirmed that 168 of these patients (93.3%) showed deficits in two or more areas of cognition (memory, visual construction, visual perception, attention, or language), consistent with the criteria for clinical diagnosis of probable AD from McKhann et al. (1984; note that these criteria were not available at the start of the original study). When also considering impairments in orientation (temporal, person, and place) and executive function, 179 of the 181 patients (98.9%) met the criteria of two or more impaired domains, demonstrating high correspondence with the McKhann et al. (1984) criteria.

An additional 54 participants included in the TalkBank release of data are described as having "Possible Alzheimer's Disease" based on the clinical team's assessments. According to McKhann et al. (1984), this diagnosis is made on the basis of dementia syndrome without other neurologic, psychiatric, or systemic explanations, which may vary in its onset, presentation, or clinical course and which may occur in the presence of a second systemic or brain disorder that is sufficient to produce dementia, but not considered to be the cause of that dementia. Here, we used the diagnosis labels ("probable" and "possible" AD vs. healthy control participants) assigned by the clinicians in the original study as the ground truth labels for analyzing participants' data in the current study. See Becker et al. (1994) for further details on participant recruitment and neuropsychological evaluations.

[2]Note that the specific details or thresholds for assessments that informed the determination of the consensus classification were not described in the original study. We report here the tests and other criteria that were administered and the behavioral/cognitive information that was considered in arriving at the consensus classification.

Participants

Using the clinical diagnoses from Becker et al. (1994), we defined two groups of participants for inclusion in the current study: *impaired participants* and *healthy controls*. In the *impaired participants* group, we included individuals who were clinically assessed as having probable or possible AD but excluded those with a main diagnosis of mild cognitive impairment (MCI), memory impairment, or other form of dementia. As some participants were clinically assessed at multiple time points and thus received multiple diagnoses over time, we based the inclusion criteria on the diagnosis that was included in the transcript header (the Codes for the Human Analysis of Transcripts [CHAT] file ID tier) to capture the diagnosis at the time point at which the speech sample was collected. Note that only the main diagnosis was entered into the CHAT file, and thus it is possible that some participants had additional diagnoses beyond possible or probable AD. The participants included in the impaired group consisted of 118 individuals with a diagnosis of probable AD and 12 individuals with a diagnosis of possible AD. The *healthy controls* group included those participants marked as control in the ID tier of the CHAT file.

In addition, we excluded participants who produced fewer than 50 words in the speech elicitation task. The number of unique participants and speech transcripts who were included in each of the analyses conducted (see Data Analysis section) are reported in Table 1.

Two analyses were conducted on the participant samples. The first analysis involved using linguistic features from the impaired participants' spontaneous speech transcripts to predict their concurrent MMSE scores; thus, only those participants and speech samples with an MMSE administered at the same time point were included, resulting in a slightly reduced number of participants included in this analysis, as shown in Table 1. Note that this resulted in the exclusion of some participants' initial assessments, which were missing recorded MMSE scores. As a result, some of the initial assessments available for analysis include MMSE scores below 10. The second analysis was binary classification, predicting whether an individual participant was in the impaired or healthy control group. For this analysis, numerical MMSE scores and demographic information were not necessary; we only required the clinical diagnosis assessed by the original study authors. We were thus able to include additional

Table 1. Number of participants included in each of the analyses reported below: (a) continuous prediction of Mini-Mental State Examination (MMSE) score (only impaired participants) and (b) binary classification of impaired versus healthy control participants.

	MMSE prediction	Binary diagnosis classification	
Variable	**Impaired participants**	**Impaired participants**	**Healthy control participants**
Number of transcripts	176	194	200
Number of individuals	127	130	93
Age: *M* ± *SD*	71.9 ± 8.58	71.7 ± 8.54	64.6 ± 8.06
Gender	38 male, 89 female	41 male, 89 female	37 male, 56 female
Education: *M* ± *SD*	12.2 ± 2.77	12.2 ± 2.76	14.0 ± 2.55
Diagnosis	115 probable AD, 12 possible AD	118 probable AD, 12 possible AD	N/A

Note. Age, gender, education, and diagnosis values are based on the first available speech transcription with an accompanying MMSE score. Diagnosis refers to the group assignment within the impaired participant sample as indicated by the original study: either probable or possible Alzheimer's disease (AD). N/A = not applicable.

participants and transcripts in this analysis as long as they had a clinical assessment, even if the time point was missing the MMSE score.

Each impaired participant completed the spontaneous speech task during at least one session and could return for follow-up sessions at approximately 1-year intervals (single session: 81 patients; two sessions: 36; three sessions: 11; four sessions: two). During each session, they completed the speech task and the MMSE. Healthy control participants returned for up to four follow-up sessions (single session: 34 participants; two sessions: 28; three sessions: 19; four sessions: seven; five sessions: five).

Healthy control participants scored numerically higher on the MMSE compared to impaired individuals during their first available assessment, with a mean score near ceiling (healthy controls: $M = 29.11$, $SD = 1.04$; impaired: $M = 18.90$, $SD = 4.99$). Healthy controls also scored higher on the MMSE when averaging the assessments across all available visits (healthy controls: $M = 29.12$; $SD = 1.13$; impaired: $M = 18.40$, $SD = 5.14$) and showed reduced declines in MMSE scores from one session to the next (healthy controls: mean difference = 0.090, $SD = 1.47$; impaired: mean difference = 3.49, $SD = 4.40$). As reported by Becker et al. (1994), the original sample of probable AD participants—which makes up the majority of our "impaired" group—showed significant deficits in several neurologic functions (e.g., olfaction, gait) and psychiatric characteristics (e.g., increased irritability, social withdrawal), relative to the healthy controls (see Becker et al., 1994, for a complete accounting of the neurological and psychiatric symptoms that were observed in the larger sample).[3] In the sample used here, the impaired participants were older on average (healthy controls: $M = 64.61$ years, $SD = 8.06$; impaired: $M = 71.70$ years, $SD = 8.54$; measured at first available assessment) and had lower levels of education (healthy controls: $M = 13.99$ years, $SD = 2.55$; impaired: $M = 12.23$ years, $SD = 2.76$). The healthy control sample had a higher proportion of males (37 out of 93 individuals, 39.78%) than did the impaired sample (41 out of 130, 31.54%). Gender was not assessed independent of sex. For more details regarding the participant make-up, see Becker et al. (1994).

In predicting MMSE scores in the sample of impaired participants, we assessed the potential of the novel, contextual linguistic features to improve detection and quantification of cognitive decline associated with dementia due to AD. The current sample is a particularly interesting test case, as it consists of participants with a very wide range of MMSE scores, yet the majority of individuals received a consensus diagnosis of probable AD. If the novel contextual features developed in the present work prove useful for prediction in the current sample, this could motivate future work that integrates these features into assessments of patients with less severe or no cognitive impairments, including MCI and subjective cognitive decline, and the prospective conversion to AD, in longitudinal studies.

It is, however, important to note that the use of MMSE scores as an outcome measure to approximate degree of cognitive impairment in dementia has limitations. While it is expected that individuals with greater cognitive decline will on average score lower on the MMSE—and indeed we observe greater reductions in MMSE scores across successive tests in AD patients than healthy controls in the current sample—its accuracy and sensitivity for detecting the conversion from MCI to dementia, and from MCI to AD, has been found to vary considerably (Arevalo-Rodriguez et al., 2015, 2021).

[3]The individual participant data for these neurological and psychiatric symptoms were not made available with the public release of the data set, so we are unable to provide the group means on these measures for our sample.

However, the ease of its use—requiring only 5–15 min to administer—has led the MMSE to be widely adopted in prior studies aiming to develop and test new screening methods for dementia. This includes several recent studies that have examined the utility of language or linguistic features for dementia and/or MCI detection (e.g., Ambrosini et al., 2019, 2024; Balagopalan et al., 2021; Beltrami et al., 2018) and multiple scientific community challenges that focused on predicting MMSE scores from linguistic features, organized by DementiaBank (Luz et al., 2020, 2021). It is also used as a standard benchmark, against which new screening tools can be evaluated (Mitchell, 2017). We thus considered it valuable to test whether the new suite of linguistic features developed in the current research improved prediction of MMSE scores in AD patients, even though it is an imperfect measure of cognitive decline or dementia severity associated with AD. If this is found to be true in the Pitt Corpus's sample, which is both uniquely large and uniquely heterogenous in terms of impairment severity, it will motivate future work that tests whether these features can predict more sensitive and comprehensive neuropsychological evaluations of AD, dementia, and MCI patients.

Procedure

As part of an extensive neuropsychological battery, participants completed a picture description task using the "Cookie Theft" image from the Boston Diagnostic Aphasia Examination (Goodglass, Kaplan, & Barresi, 2001). Participants were shown the image and asked to describe everything in the depicted scene while their spoken responses were recorded. The recordings were then transcribed into text files using the CHAT protocol (MacWhinney, 2000). We only included transcripts from impaired individuals and healthy controls, which contained at least 50 words in the description.

Linguistic Features

Seventeen linguistic features were calculated on each transcript using custom Python (Version 3.10.8) scripts (see Table 2 for a description of the features used in the present work). A subset of the features was selected based on prior clinical research demonstrating a relationship between these features of spontaneous speech and AD. Specifically, past work has associated AD with changes that include increased production of empty or indefinite words (Croisile et al., 1996; Ehrlich, Obler, & Clark, 1997; Hier et al., 1985; Nicholas et al., 1985), higher frequency words (Fraser et al., 2015; Kavé & Goral, 2018), and function words (Almor, Kempler, MacDonald, Andersen, & Tyler, 1999; Fraser et al., 2015; Hier et al., 1985; Kavé & Goral, 2016). AD patients also show reduced production of content words (Ahmed et al., 2013; Kavé & Goral, 2016) and definite references to objects (Feyereisen et al., 2007), as well as more frequent production of irrelevant information (Carlomagno et al., 2005; Croisile, et al., 1996; Forbes et al., 2002; Murray, 2010)

Table 2. List of linguistic features and their corresponding categories.

Category	Feature	Explanation
Word-level	Total words	Total count of words spoken, including real words, nonwords, and partial words
	Fillers†	Count of filler words (e.g., "um," "uh")
	Empty words†	Count of empty words (e.g., "stuff," "thing")
	Definite articles†	Count of the definite article "the"
	Indefinite articles†	Count of indefinite articles "a" and "an"
	Pronouns†	Count of pronouns
	Nouns†	Count of nouns
	Verbs†	Count of verbs
	Content words†	Count of all real words that are not function words
	Lexical frequency	Median of the log of the frequency of all real words
	Type–token ratio	Ratio of unique words to total words spoken
	Honoré's statistic	Measure of lexical richness based on the number of words produced exactly once
Contextual: Surprisal	Median surprisal	Median surprisal calculated across a speech transcript
	IQR of surprisal	Interquartile range of surprisal calculated across a speech transcript
Contextual: Fillers	Content word frequency after filler	Median frequency of the next content word following a filler word
	Distance to next content word	Median distance, in words and nonwords, from a filler word to the next content word
	Surprisal after filler	Median surprisal of the next real word after a filler

Note. Features marked with † were normalized by the length of the transcript. IQR = interquartile range.

and/or production of fewer words that are relevant to the discourse (Croisile et al., 1996; Feyereisen et al., 2007; Fraser et al., 2015; Nicholas et al., 1985). Finally, AD patients also show an overall reduction in lexical diversity, that is, producing fewer unique words and more repetition of the same words in their speech (Bucks et al., 2000; Hier et al., 1985; Mueller et al., 2016).

On the basis of these findings, we selected 12 discourse features that are defined at the individual word level and have attested relationships with AD from previous work, which we refer to as "word-level features" for convenience. These features were derived from the production of isolated words (i.e., not taking into account the word's location within the language context) and included the total number of words, median lexical frequency, type–token ratio (the ratio of unique words to total words), and Honoré's statistic (a measure of lexical richness based on the number of words produced exactly once), as well as counts of various lexical categories: filler words (e.g., *um*, *uh*), empty words (e.g., *stuff*, *thing, place*), definite and indefinite articles, pronouns, nouns, verbs, and content words. For the latter count features, the feature was normalized to the length of the transcript by dividing the raw count of that feature by the square root of the total number of words in the transcript. Part of speech counts were determined using the Natural Language Toolkit in Python (S. Bird et al., 2009) and the Penn Treebank tag set (Marcus et al., 1993). Lexical frequency was calculated using the Switchboard and Fisher corpora, a collection of spoken telephone conversations consisting of 24 million words and 1,975 hr of speech (Cieri et al., 2004, 2005; Godfrey & Holliman, 1993).

The remaining five linguistic features were novel *context features*, meaning they were calculated on each word in the transcript based on the larger linguistic context in which it occurred. Three context features were based on lexical surprisal (also known as Shannon information content; Shannon, 1948), defined as the negative logarithm of the probability of a word within its context, in other words, how unexpected a particular word is given the preceding linguistic context. Word surprisal values extracted from large language models have been shown to correlate with behavioral measures of language processing difficulty (e.g., Balling & Baayen, 2012; Smith & Levy, 2013) and human brain activity during language comprehension (e.g., Brennan & Hale, 2019; Frank et al., 2015). In speech production, these surprisal values can thus operationalize the difficulty a speaker experiences while undergoing lexical selection and planning. They can also be viewed as a measure of how far a speaker deviates from the language patterns and transitional probabilities between words that are "expected" based on the substantial corpus (approximately 40 GB) of training text data.

The probability of each word given the preceding words that were produced in the speech stream was computed using the pretrained large language model GPT-2 (Radford et al., 2019), which was trained on a corpus of over 8 million documents. We selected GPT-2 to compute word-by-word surprisal for two reasons: First, it had the desirable property that it was trained using unidirectional, next-word-prediction, meaning that each target word is predicted on the basis of only the previous words in a text, rather than bidirectional prediction that considers words that both precede and follow a target word, as is the case in other computational language models (e.g., the BERT model; Devlin et al., 2018); thus, GPT-2's calculations of surprisal are aligned with how humans process language. Second, several results have demonstrated that predictions or contextualized word embeddings extracted from GPT-2 correlate with human ratings of word surprisal or neural responses during naturalistic listening (Cauchteaux et al. 2021; Goldstein et al., 2020, 2024).

We retrieved the GPT-2 model from the huggingface online repository (http://www.huggingface.co) and used the *minicons* (Misra, 2022) and *transformers* (Wolf et al., 2020) Python libraries to convert each transcript to a series of word tokens. The conditional probability of each token was calculated using a sliding window of the previous 12 tokens and converted to surprisal by taking the negative log of the conditional probability. At the start of a transcript (beginning from the second token), the context window was incrementally increased to 12.

After computing lexical surprisal for each word in the transcript, we calculated three related features: (a) the median surprisal across all tokens in a transcript, (b) the interquartile range of surprisal within a transcript (capturing the spread), and (c) the median surprisal of words immediately following a filler word (e.g., "the boy is um, falling," surprisal of *falling*). In the latter case, surprisal was calculated after removing the fillers from the 12-token window over which the word's conditional probability was calculated (i.e., the probability was calculated on the linguistic context as if that filler word, as well as any others in the preceding context window, had not occurred). The final two context features were also defined based on properties of utterances that followed the appearance of a filler word. These were (d) the median distance from a filler to the next content word (e.g., "the boy is grabbing um at the plate," distance = 3 words) and (e) the median lexical frequency of the next content word after a filler (e.g., frequency of *plate* in the previous example).

In addition to these 17 linguistic features (12 word-level from prior literature and five novel context features),

three demographic properties were used as predictors in the linear regression: the participant's age (years), sex (male/female), and education (years). Thus, there were a total of 20 predictors in the statistical model: 17 linguistic features and three demographic features.

Data Analysis

Prediction of Impaired Participants' MMSE Scores From Linguistic Features

To assess whether the automatically extracted linguistic features could predict impaired participants' continuous neuropsychological test scores, we built a multiple linear regression model in R (Version 4.2.2; R Core Team, 2021) to predict each participant's MMSE score as a function of all linguistic features entered jointly as predictors. For each analysis, two multivariate regression models were constructed: (a) a baseline model, which included only the three demographic variables (age, sex, years of education) as predictors, and (b) a test model, which included the three demographic variables plus the 17 linguistic variables as predictors. In both cases, the outcome variable was MMSE score. We used model comparison to test whether the linguistic features added explanatory power for MMSE score relative to demographic variables alone.

To test whether our novel context features explained variance beyond the individual word-level features, we also compared a model containing only the word-level and demographic properties as predictors against one additionally containing the context features. This was conducted first using the full set of five context features and then using two subsets: those primarily related to filler usage (distance to next content word following a filler, median frequency of the content word, and surprisal of the next word following a filler) and those remaining context features derived from surprisal (median surprisal, interquartile range of surprisal; see Table 2 for feature categories). All analyses were conducted using two sets of transcripts from impaired participants: (a) restricted to the first available assessment for each participant (n = 127 transcripts) and (b) the complete set of assessments, including all follow-ups (n = 176 transcripts). (As noted in Participants section, the first available assessment with an MMSE score was not necessarily the initial assessment.) See Table 3, rows 1 and 2, for the profile of participants included in these analyses.

Classification of Impaired Versus Healthy Control Participants

To test whether the complete set of computed linguistic features could accurately classify an individual as impaired or a healthy control, we performed a fivefold stratified cross-validation classification analysis in Python using the *scikit-learn* library (Buitinck et al., 2013; Pedregosa et al., 2011). The complete sample (i.e., all transcripts from the healthy controls and impaired patients) was divided into five folds, with numbers of each diagnosis group proportional to the full sample. On each iteration of model training, a logistic regression classifier with L2 regularization was trained on four of the five folds using nested cross-validation to select an optimal regularization weight parameter (from five values on a logarithmic scale between $1e^{-4}$ and $1e^{4}$). The training data were first standardized so that each predictor and the outcome variable had a mean of zero and unit variance. The model was then trained on the four training folds to optimize its regularization parameter for classification performance, based on the area under the receiver

Table 3. Summary of analyses, participant counts, and Mini-Mental State Examination (MMSE) scores.

Analysis group	Dependent variable	No. of speech transcripts	Mean MMSE	MMSE range
Impaired participants, all sessions	MMSE	Impaired: 176	Impaired: 18.40 (*SD* = 5.14)	Impaired: 1–30
Impaired participants, 1st sessions only	MMSE	Impaired: 127	Impaired: 18.90 (*SD* = 4.99)	Impaired: 3–30
Impaired & healthy participants, all sessions	Binary group classification	Impaired: 194 Control: 200	Impaired: 18.40 (*SD* = 5.14) Control: 29.12 (*SD* = 1.13)	Impaired: 1–30 Control: 24–30
Impaired & healthy participants, 1st sessions only	Binary group classification	Impaired: 130 Control: 93	Impaired: 18.90 (*SD* = 4.99) Control: 29.11 (*SD* = 1.04)	Impaired: 3–30 Control: 26–30

Note. Two analytical approaches were employed. First, we attempted to predict impaired participants' MMSE scores based on linguistic features. This was done separately using data from all sessions (first row) and data from only the first available sessions (second row). The second analysis approach was performing binary classification of healthy controls versus impaired participants based on their linguistic features. This was again separately conducted using data from all sessions (third row) and data from only the first available sessions (fourth row). Note that some patients were classified as impaired but did not have MMSE scores recorded in the data set; as a result, the group classification analyses have a slightly higher *N* than the continuous MMSE prediction analyses. The mean and standard deviation of MMSE scores in these groups were calculated using those participants with recorded scores.

operator characteristic (ROC) curve: an aggregate measure of model performance across decision boundary thresholds. The left-out fold—the test set—was then used to assess the model's performance on unseen data. This process was repeated, leaving one fifth of the data out each time, and the mean area under the ROC curve calculated by averaging across all five folds.

If linguistic features provide sufficient information to distinguish between individuals with and without dementia, the average performance of the model should be statistically greater than that expected by chance alone. To assess this, we permuted the transcript labels (impaired vs. healthy control) 1,000 times, so that the diagnosis labels were randomly assigned to each participant, and repeated the entire training and testing procedure, recording the model's average performance across the five folds each time. We then compared the true performance to this surrogate null distribution to estimate a p value. If the classification accuracy of the model created with the true diagnosis labels was better than 95% of those created using the permutated labels—corresponding to a p value of .05—it was considered statistically significant. As with the prediction of MMSE, the classification analysis was conducted using two samples: (a) only the first available assessments and (b) the full set of assessments, including all follow-ups. See Table 3, rows 3 and 4, for the profile of participants included in these analyses.

Post Hoc Exploratory Analyses

We performed two post hoc exploratory analyses. First, to examine the relationship between individual linguistic features and cognitive impairment, simple regressions between each linguistic feature and MMSE score were performed, with p values corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate method. Second, we performed principal components analysis on the feature values extracted from the transcripts of dementia patients to examine the latent structure underlying the complete set.

Results

Prediction of MMSE Scores of Impaired Participants From Linguistic Features

Several multiple linear regression models were constructed, using the 12 individual word-level features, five novel context features, and three demographic properties (age, sex, and education), entered jointly as predictors, to predict MMSE scores. For each analysis, we report results from (a) only the first available assessments and (b) all assessments, including first available and follow-ups, without modeling repeated measures (i.e., each transcript is considered independently). See Table 4 for a summary of the statistical results from the models discussed in this section.

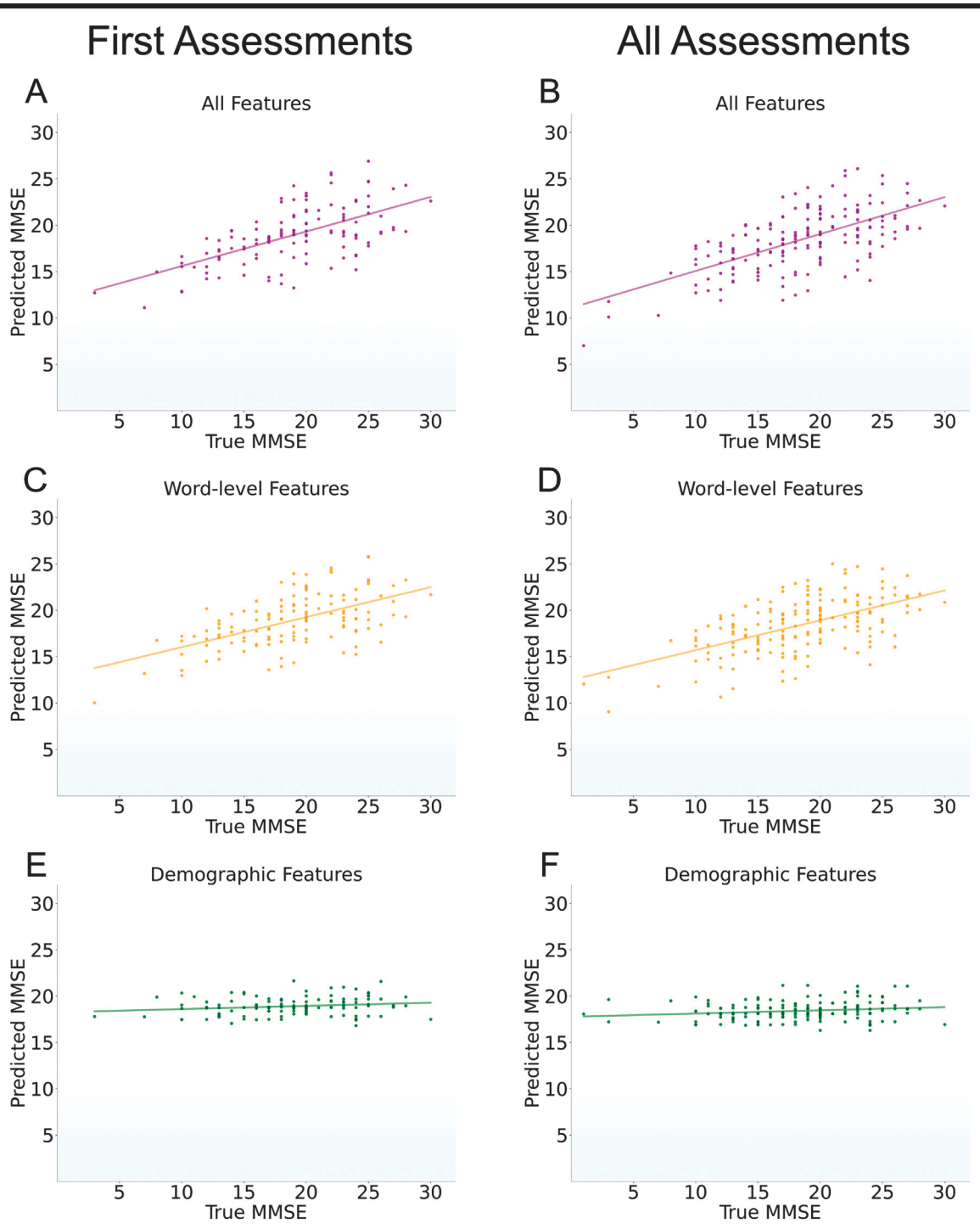
As shown in Figures 1A and 1B, the complete model containing all linguistic and demographic features as predictors significantly predicted MMSE scores in impaired patients both when considering patients' first assessments (n = 127; adjusted R^2 = .254, $F(20, 106)$ = 3.149, $p < .001$; see Figure 1A) and also when considering first assessments and all follow-ups (n = 176; adjusted

Table 4. Summary of statistical results from each of the multiple linear regressions conducted to predict patients' Mini-Mental State Examination scores.

Model comparison	First assessments	All assessments
M1: Demographic	M1 adjusted R^2 = .011	M1 adjusted R^2 = .018
M2: Demographic + word-level + context	M2 adjusted R^2 = .254	M2 adjusted R^2 = .320
M1 vs. M2	$F(17, 106) = 3.362$, $p < .001$	$F(17, 155) = 5.498$, $p < .001$
M1: Demographic + word-level	M1 adjusted R^2 = .232	M1 adjusted R^2 =. 259
M2: Demographic + word-level + context	M2 adjusted R^2 = .254	M2 adjusted R^2 =. 320
M1 vs. M2	$F(5, 106) = 1.660$, $p = .151$	$F(5, 155) = 3.899$, $p = .002$
M1: Demographic + word-level	M1 adjusted R^2 = .232	M1 adjusted R^2 = .259
M2: Demographic + word-level + surprisal	M2 adjusted R^2 = .264	M2 adjusted R^2 = .318
M1 vs. M2	$F(2, 109) = 3.420$, $p = .036$	$F(2, 158) = 7.94$, $p < .001$
M1: Demographic + word-level	M1 adjusted R^2 = .232	M1 adjusted R^2 = .259
M2: Demographic + word-level + filler	M2 adjusted R^2 = .223	M2 adjusted R^2 = .260
M1 vs. M2	$F(3, 108) = 0.559$, $p = .643$	$F(3, 157) = 1.073$, $p = .362$

Note. Each row lists the base model (M1) and comparison model (M2), showing which predictors were used in that regression. The M1 and M2 in each row were statistically compared to determine whether the predictors added in M2 explained additional variance over and above those included in the baseline M1. The F and p values reported in the second and third columns are those from the model comparison. Each M1 vs. M2 row refers to the two models defined above it.

Figure 1. Regression results showing the relationship between true Mini-Mental State Examination (MMSE) scores and predicted MMSE scores, for impaired participants. The left column shows regressions computed using only the first available assessments, while the right column shows regressions computed using all available assessments. (A, B) The predicted MMSE scores from a model containing all linguistic features, including word-level and the novel context features, as well as demographic properties. (C, D) Predicted scores using only the word-level features. (E, F) The predicted scores using only the demographic features (sex, age, years of education). For illustration purposes, each plot includes a line of best fit based on the true and predicted MMSE scores.



R^2 = .320, $F(20, 155) = 5.119$, $p < .001$; see Figure 1B). Importantly, in both cases, the full model including the linguistic variables explained significantly more variance than did the baseline model with only demographic properties as predictors (first visits: $F(17, 106) = 3.362$, $p < .001$; all visits: $F(17, 155) = 5.498$, $p < .001$; models with only demographic properties shown in Figures 1E and 1F). This establishes that our computed set of linguistic features can successfully explain a meaningful amount of variance in the degree of cognitive impairment, as operationalized by MMSE scores.

We next asked whether the novel context features added a significant amount of explained variance relative to the set of demographic and word-level predictors. To do so, we compared a model containing the word-level and demographic predictors to one containing word-level, demographic, and, additionally, context features as predictors. When considering only participants' first assessment, adding the five context features improved the amount of explained variance numerically (adjusted R^2 = .254 [demographic + word-level + context features] vs. .232 [demographic + word-level features]), but this was not a statistically significant difference, $F(5, 106) = 1.660$, $p = .151$. When including all assessments, the five context features significantly improved MMSE prediction (adjusted R^2 = .320 [demographic + word-level + context features] vs. .259 [demographic + word-level features]); $F(5, 155) = 3.899$, $p = .002$, demonstrating that the set of context features captured a significant portion of variance in MMSE scores and added explanatory power.

Since the context properties can be further divided into two categories themselves (surprisal-related and filler-related; see Table 2), we also examined the unique contribution of each category over and above the demographic and word-level variables as predictors. Adding median surprisal and the interquartile range of surprisal to models that contained only the word-level and demographic properties resulted in significantly better prediction of MMSE scores. This improvement was statistically significant when considering only the first assessments (adjusted R^2 = .232 [demographic + word-level features] vs. .264 [demographic + word-level + surprisal features]; $F(2, 109) = 3.420$, $p = .036$), as well as when considering all of the assessments (adjusted R^2 = .259 [demographic + word-level features] vs. .318 [demographic + word-level + surprisal features]; $F(2, 158) = 7.94$, $p < .001$). Adding only the filler-related features, however, did not significantly improve prediction of MMSE score relative to models that already contained the word-level and demographic properties in either case (first visits: adjusted R^2 = .232 [demographic + word-level features] vs. .223 [demographic + word-level + filler features]; $F(3, 108) = 0.559$, $p = .643$; all visits: adjusted R^2 = .259 [demographic + word-level features] vs. .260 [demographic + word-level + filler features]; $F(3, 157) = 1.073$, $p = .362$). This suggests that while the surprisal of a speaker's words within the larger discourse context captures meaningful variance in level of cognitive impairment, in both initial and follow-up assessments, the filler-related features do not.

Classification of Impaired Versus Healthy Control Participants

In a second analytical approach, we assessed whether the set of linguistic features could be used to accurately classify individuals as healthy controls versus impaired individuals. A fivefold stratified cross-validation was used to train regularized logistic regression models to predict the category of each transcript (impaired/control) based on the automatically extracted linguistic features, excluding the demographic properties. Each model's performance was evaluated by calculating the area under the ROC curve in left-out data.

The mean area under the ROC curve across all folds was 0.871 (SD = 0.035), with a mean classification accuracy of 79.2% (SD = 3.29%). A permutation test using 1,000 random permutations of the category labels demonstrated that this classification performance was significantly greater than what would be expected by chance alone ($p < .001$). For comparison, a model that classified every transcript as a healthy participant (the majority class) would achieve an accuracy of approximately 51%, based solely on the proportion of each class (200 out of 394 transcripts were from impaired participants). Slightly lower but similar model performance was found when using only the transcripts from the initial visit, where the mean area under the ROC curve was 0.829 (SD = 0.032) and the mean accuracy was 75.4% (SD = 8.00%, $p < .001$). This was significantly above the chance level of 58.3% if classifying every transcript as impaired based on the proportion of impaired individuals (130 out of 223 transcripts).

In addition to evaluating the models based on accuracy and ROC curves, we also explicitly examined the model's performance when prioritizing correct detection of impairment versus correct rejection of healthy individuals. To do so, we examined precision and recall when we shifted the model's decision boundary, which is the amount of evidence necessary for the model to classify an individual as impaired, across a range from 0.35 to 0.65 in steps of 0.01. Precision refers to the ratio of true positives to the total number of predicted positives (i.e., true positives + false positives) and thus captures what proportion of the participants who were predicted to be impaired that actually were impaired. Recall refers to the ratio of correctly predicted (true) positives to the total number of

actual positives (including those that the model incorrectly labeled as negative) and thus captures the model's ability to detect impairment (i.e., the proportion of impaired participants who were predicted to be impaired). For example, if the model predicted all participants to be impaired, it would have low precision (because many of the predicted impaired participants would actually be healthy) but high recall (because all impaired participants would be correctly predicted as impaired). For ease of interpretation, we converted both ratios to percentages.

At the default decision boundary of 0.50, the model's mean precision and recall across the five training folds were 78.9% and 78.7%, respectively. As expected, the model's recall increased as we decreased the decision boundary; see Figure 2. At a threshold of 0.35, the classification demonstrated a mean recall of 87.6% (i.e., correct detection of impairment in 170 of 194 impaired individuals in the current data set) with a precision of approximately 73.0%. This demonstrates that the current approach could be tailored to increase sensitivity (i.e., increasing the recall score), if the failure to detect dementia in an impaired person may be considered of greater consequence than a false positive that could be cleared in follow-up consultation with a medical professional.

Post Hoc Exploratory Analyses

To explore the relationship between individual linguistic features and cognitive impairment, Pearson correlation coefficients were calculated between each of the features and MMSE score, using the data from all impaired participant transcripts (n = 176). All p values were corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate method (Benjamini & Hochberg, 1995). Two individual word-level features showed significant individual correlations with MMSE score at an adjusted threshold of $p < .01$: median lexical frequency ($r = -.45$, $p < .001$) and proportion of definite articles ($r = .32$, $p < .001$). One word-level feature implicated in past studies of speech in dementia was significant at a threshold of $p < .05$ after correction: The usage of nouns ($r = .24$, $p = .020$) and the usage of empty words were marginally significant ($r = -.21$, $p = .053$). Two novel context features were individually correlated with MMSE score: median word surprisal ($r = -.290$, $p = .002$) and the frequency of the next content word after a filler ($r = -.240$, $p = .020$). These individual correlations suggest that individuals with greater impairment (i.e., lower MMSE scores) produce more frequent yet more surprising words, make greater usage of empty words, and use nouns and definite articles less frequently (see Figure 3).

Figure 2. Classification of impaired versus healthy control individuals. Model precision and recall are shown, averaged across five folds of data, as a function of the model's threshold for classifying an individual as impaired. A lower decision threshold means less evidence is required to classify a participant as impaired.

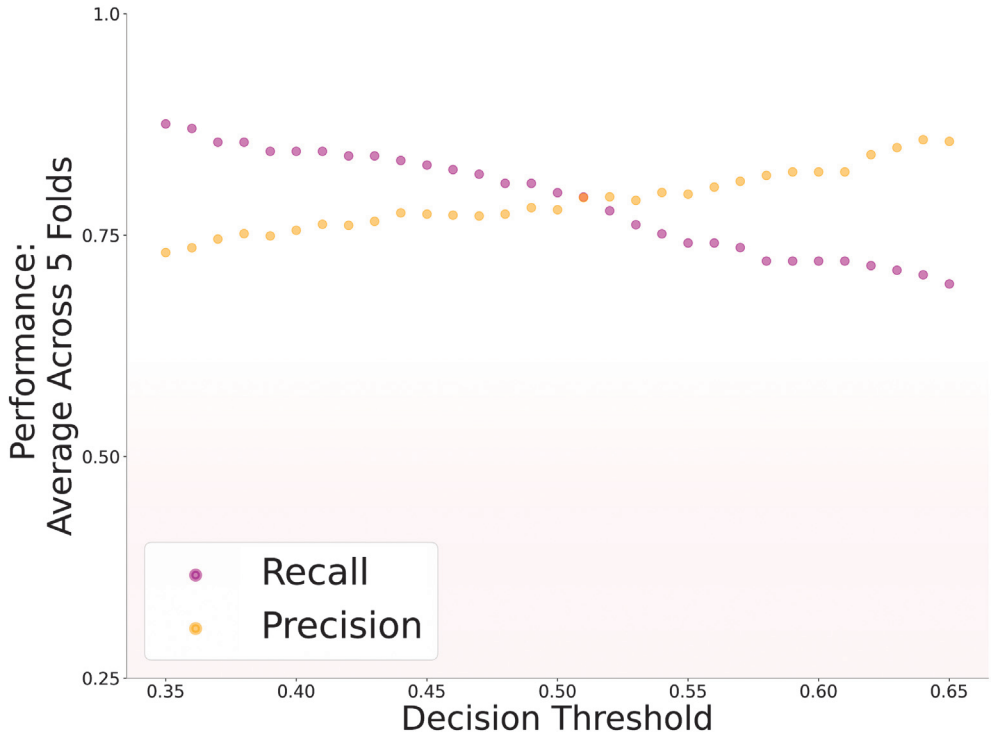
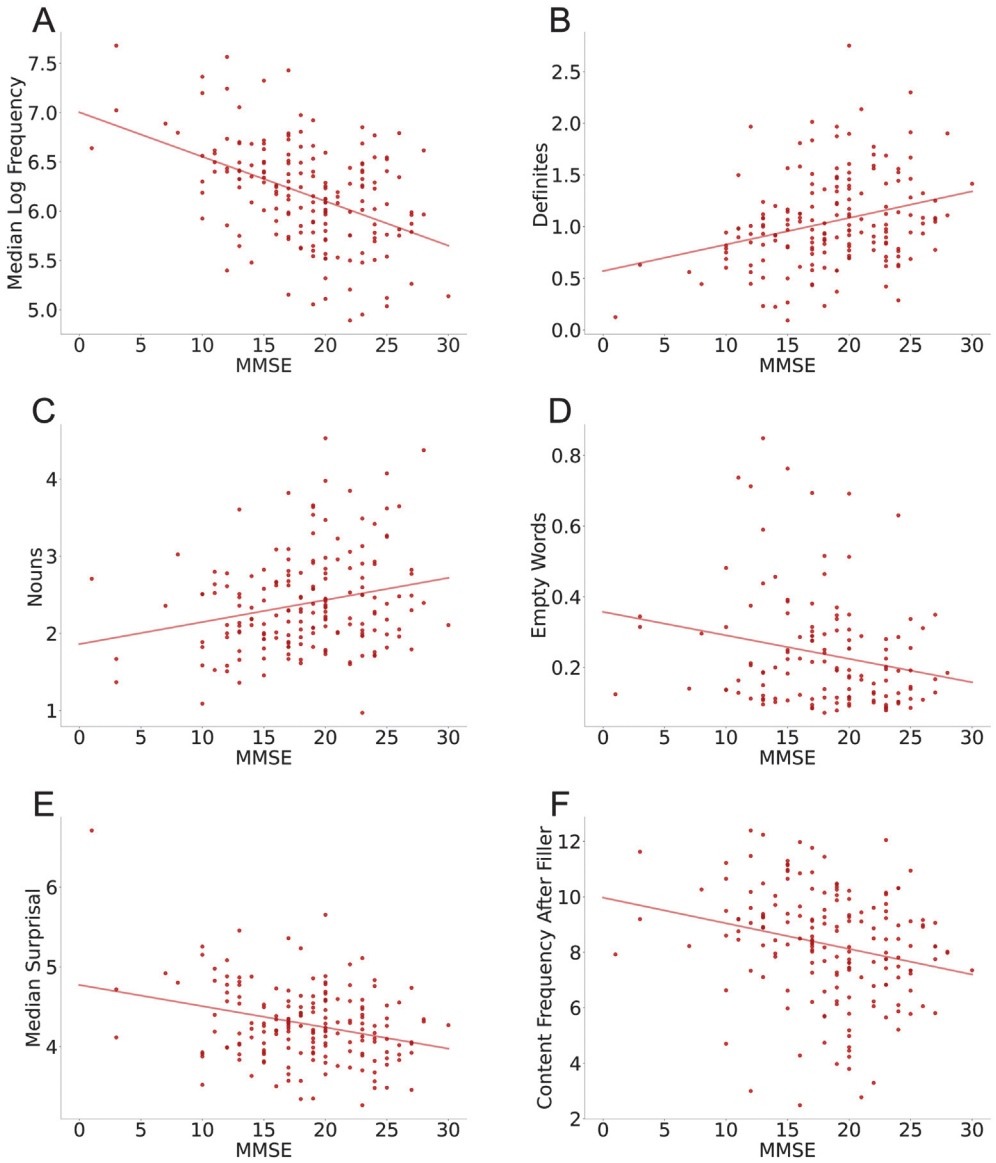


Figure 3. Individual linguistic features that significantly predicted Mini-Mental State Examination (MMSE) scores. Significant individual Pearson correlations were observed between MMSE score and the individual word-level features: (A) median log frequency, (B) use of definite articles, and (C) usage of nouns. A marginally significant relationship was found between usage of (D) empty words and MMSE. Two novel context features showed significant correlations with MMSE score: (E) median word surprisal and (F) the mean frequency of content words that followed a filler. For illustration purposes, each plot includes a line of best fit based on the feature and MMSE score.



Lastly, we used principal components analysis (PCA) to examine the latent structure underlying the set of linguistic features automatically extracted from the impaired patients' data. Three components were found to together explain approximately 53% of the variance in the set of features (26.0%, 16.6%, and 10.3%, respectively), with subsequent components each explaining approximately 7% or less. As shown in Table 5, the first component weighted heavily on a set of features that included a higher total number of words, lower type–token ratio, and greater use of content words and verbs. Component 2 was associated with lower overall lexical frequency and lower lexical frequency after filler words, as well as fewer empty words, more definite articles and nouns, but (to lesser degree) fewer pronouns. Component 3 was unique in its clear weighting on context measures, with higher lexical surprisal (median surprisal and interquartile range), along with higher lexical richness measured by Honoré's statistic and reduced usage of fillers. Component 1 was marginally significantly correlated with MMSE score ($\rho = 0.146$, $p = .053$), and Components 2 and 3 were both significantly correlated with MMSE (Component 2: $\rho = 0.382$, $p <$

Table 5. The set of linguistic features and loadings on each of the first three principal components (PCs).

Feature class	Feature name	PC1	PC2	PC3
Word-level	Total words	**0.448**	0.028	0.070
	Fillers†	0.150	–0.034	**0.391**
	Empty words†	0.176	**–0.309**	0.087
	Definite articles†	0.153	**0.446**	–0.163
	Indefinite articles†	0.147	–0.019	0.027
	Pronouns†	0.316	–0.293	0.064
	Nouns†	0.258	**0.393**	0.189
	Verbs†	**0.364**	–0.189	0.012
	Content words†	**0.410**	0.046	0.197
	Lexical frequency	–0.019	**–0.500**	0.040
	Type–token ratio	**–0.378**	–0.097	0.240
	Honoré's statistic	–0.253	0.020	**0.364**
Context: Fillers	Content word frequency after filler	0.074	**–0.316**	–0.050
	Distance to next content word	–0.048	–0.107	0.228
	Surprisal after filler	–0.027	0.077	0.051
Context: Surprisal	Median surprisal	–0.130	0.030	**0.562**
	IQR of surprisal	0.030	0.077	**0.404**

Note. Features loading above an absolute value of 0.300 is marked in bolded font. Features marked with † are those normalized by the square root of the total number of words. IQR = interquartile range.

.001; Component 3: $\rho = -0.177$, $p = .019$). Thus, Component 1 can be interpreted as demonstrating that people with greater impairment produce speech with fewer content-bearing words. Component 2 captures the pattern that people with greater impairment produce less specific language: fewer specific nouns and definite articles and more pronouns and empty words, as well as higher frequency words (which tend to be less specific and precise). Finally, Component 3 can be interpreted as showing that participants with greater impairment produce words that are more surprising (i.e., less predictable) within the context and more filler words.

Discussion

Linguistic changes in AD and other forms of dementia have been well documented (e.g., Ahmed et al., 2013; Mueller et al., 2016; Mueller, Hermann, et al., 2018), indicating the potential of language-based screening methods for improved detection of dementia. Here, we examined two new sets of features, which take into account the discourse context that words are produced in, for predicting MMSE scores as a proxy of dementia severity in patients with probable or possible AD. These novel features sets were (a) properties of words that immediately followed the occurrence of a filler (e.g., *he is um reaching*) and (b) the surprisal (i.e., unexpectedness/predictability) of individual words given those that preceded them.

We began by demonstrating that an encompassing set of linguistic features—including 12 features selected from prior findings as well as five novel context features—could jointly predict MMSE scores in AD patients, performing better than models that used demographic properties alone. We demonstrated these linguistic features' utility both in inferring a continuous score that is expected to decrease, on average, with participants' dementia severity and in performing binary classification to detect whether a participant was in the healthy or impaired group. Next, we showed that novel context properties explained a significant proportion of variance in MMSE scores, suggesting that inclusion of these linguistic properties could improve speech-based screening methods for dementia due to AD, including earlier in disease progression.

We specifically examined the relevance of two subgroups of the novel context features: those characterizing the participant's use of fillers (surprisal and frequency of words following a filler, and distance to the next content word) and those characterizing the surprisal of the participant's lexical production (median surprisal across all words, interquartile range of surprisal). While the new filler-related features did not explain a significant proportion of variance in MMSE score beyond what was captured by the previously attested word-level properties, the surprisal features significantly improved model performance. Moreover, median surprisal was individually negatively correlated with MMSE score, suggesting that with greater cognitive impairment due to AD, individuals tended to deviate further from the language model's statistical predictions of what words would be said next. MMSE scores also showed positive relationships with the usage of nouns and definite article determiners, and

negative relationships with empty words and median lexical frequency.

These findings provide novel evidence that the incremental surprisal of Alzheimer's patients' words in connected speech explains unique variance in cognitive impairment (i.e., degree of dementia severity), as can be approximated by MMSE scores, and beyond that explained by previously examined properties (e.g., word frequency, number of nouns) that have been the focus of past work. That is, the language produced by people who are more impaired is less predictable based on the preceding linguistic context. This suggests that with greater cognitive decline, individuals with AD produce more empty words, fewer nouns and definite articles and tend to use more common yet also more surprising words in picture descriptions. Together, this constellation of impairments in language production suggests that AD leads to language with reduced semantic specificity, as these linguistic deficits all result in the production of less specific and content-heavy words (i.e., emptier words, more common words, words that fit less well within the ongoing context, and fewer nouns and definite articles to pick out specific as opposed to more generic referents).

The Relationship Between Frequency, Surprisal, and Cognitive Impairment

These results demonstrate a curious pattern regarding the relevance of both surprisal and frequency in predicting MMSE scores. Several previous studies have reported that the median or mean lexical frequency of words produced by patients with AD or other forms of dementia or MCI is higher than that observed in the speech of cognitively unimpaired individuals (H. Bird et al., 2000; Fraser et al., 2015; Kavé & Goral, 2016; Ostrand & Gunstad, 2021), which was also shown in the current results. Past work has theorized that this is due to an underlying impairment in lexical access—the process of retrieving words from memory to produce them in speech (Burke & Shafto, 2008)—that may be attributed to degradation of medial temporal lobe and temporoparietal areas in dementia (Whitwell et al., 2007). This impairment is theorized to result in higher average word frequency in patients' connected speech because more common words are less effortful to retrieve from one's mental lexicon. Here, we provide what we believe is the first evidence that a related variable—lexical surprisal—captures additional variability in cognitive impairment in individuals with AD.

Devoid of sentential context, the frequency of a word should be inversely proportional to its surprisal, as more common words are less surprising on their own (i.e., they carry fewer bits of information). A natural inference may thus be that individuals with greater cognitive decline would show reduced lexical surprisal in their speech, due to the increased lexical frequency of the words they produce. However, in the current work, we found negative relationships between MMSE scores and both frequency and surprisal; that is, language production from individuals with lower MMSE scores tended to display higher median frequency (as in prior studies) and yet also higher median surprisal. On the other hand, frequency and surprisal were not significantly correlated across patient transcripts ($r = -.045$, $p = .538$) and the relationship between surprisal and MMSE score remained significant after partialing out the relationship between MMSE and frequency (partial correlation, $p < .001$). Moreover, the two variables dissociated on orthogonal dimensions in the PCA decomposition (see Table 5). Together, the current findings suggest that lexical surprisal predicts degree of cognitive impairment (as operationalized by MMSE score) separately from the relationship with lexical frequency found in previous work.

To better understand what may be driving the relevance of lexical surprisal, we examined the points in individual transcripts where surprisal was higher than the 75th percentile. A clear pattern emerged: Words with particularly high surprisal were often incomplete, corrected, or repeated (e.g., "*There's a lit-, a girl, young girl . . .* "; underlined words are those with high surprisal). Qualitative inspections also suggested that high surprisal words tended to appear after word choice errors, as in " *. . . they're grading, uh they, they are going to um get some cookies . . . ".* In this example, the speaker mistakenly describes children who are *grading*, despite no depiction of this in the image. Both *grading* and *cookies* have high estimated surprisal, presumably because the former is a mistake and unexpected given the statistics of language usage (children do not often grade), while the latter is unexpected given the previous sentential context that describes grading (a less surprising noun in this context may be *pencils*).

This example highlights that our operational definition of surprisal may be capturing the contextual relevance of not only one's word choices but also an individual's ability to produce coherent, error-free speech. This effect may result from the fact that the GPT-2 language model was trained primarily on written language, from over 8 million online documents. Because one can revise written language before publishing it, these documents likely contain fewer word choice errors and repetitions than is typical of spoken language. Indeed, past research has found that individuals with a heightened risk for AD or with very early mild cognitive impairment show reductions in speech fluency in their connected speech (Mueller, Koscik, Hermann, et al., 2018), which might (partially) account for increased linguistic surprisal in discourse production.

An important related measure, borrowed from aphasia research, which also captures the typicality of lexical production in connected speech, is known as core lexicon analysis. For a given speech elicitation task—most commonly a picture description or procedural narrative—core lexicon items are determined by collecting discourse samples from healthy control participants and collating the words that are produced across participants (Dalton et al., 2020; Dalton & Richardson, 2015). Thus, core lexicon analysis provides a set of expected words for a particular speech elicitation task, and a patient's transcript can be compared against this normed list to provide a score of how "typical" their word production is. The linguistic surprisal metric in the current work is a way of operationalizing a similar underlying measure as core lexicon—that of the expectedness or typicality of the words produced by the participant—but has the advantage that its use is not restricted to the small number of constrained speech tasks for which core lexicon standards have been developed. It also does not require the collection and hand-coding of an additional set of norming data from participants performing the same speech tasks. On the other hand, core lexicon analyses use a corpus built from connected speech from healthy participants for a particular picture description and thus may be more uniquely targeted to the individual elicitation task by providing a measure of the typicality of an individual's words in a very precise context, whereas our surprisal metric uses a corpus built from vast troves of generic language on the internet.

An interesting direction for future work would be to examine the degree to which these two approaches—core lexicon analysis and computationally derived linguistic surprisal—may be capturing the same underlying cognitive and linguistic changes in dementia. Initial research has shown that patients with MCI and/or AD display deficits in producing core lexicon items in picture description tasks as well (Chen et al. 2025; Kintz et al. 2024), adding support to the relevance of surprisal as a linguistic property for detecting MCI and AD.

Comparisons to Past Work

Previous attempts to use linguistic features to classify individuals as patients with dementia or healthy controls have demonstrated accuracies that range from approximately 75% up to 90%–95% correct (e.g., Bucks et al., 2000; Guinn & Habash, 2012; Jarrold et al., 2014; Meilán et al., 2014; Thomas et al., 2005). Two studies have done so using the same DementiaBank Pitt data set that we employed here and, thus, the same distinction between AD patients and healthy controls (although participant exclusion may have varied). Orimaye et al. (2014) compared the performance of various machine learning algorithms tasked with classifying the impaired and healthy control participants based on a relatively restricted set of features that included syntactic properties (e.g., number of predicates, number of coordinated sentences) and select lexical properties (e.g., number of function words, unique words, repetitions, and morphemes). They found that support vector machines trained on these features could classify the individuals with an F-measure score (the harmonic mean of recall and precision) of 74%. Fraser et al. (2015), on the other hand, considered 370 features derived from the speech transcripts and the audio recordings of the participants' picture descriptions and assessed the performance of logistic regression models that included increasing numbers of those features, from one to 370. Their peak classification accuracy was approximately 82%, which was observed when using 35 of the features.

Our results are similar to those obtained in these prior studies, with a mean area under the ROC curve of 0.871 and a mean accuracy of 79.2% when using a standard decision boundary. Our observed area under the ROC curve is also within a reasonable range of that reported for the MMSE itself in detection of cognitive impairment and dementia (0.890; for comparison, the Modified Mini-Mental State Exam has a reported area under the ROC curve of 0.930; McDowell et al., 1997), but this should be interpreted with caution as the evaluation samples may differ in the severity and progression of disease. The small reduction in accuracy that we observe, relative to the results of Fraser et al. (2015), may be attributable to the data-driven feature selection and iteration process that those authors adopted, which enabled them to identify the most informative number and set of features from a much larger number of candidates. Here, we took a hypothesis-driven approach for choosing a smaller set of features, using only the speech transcripts, and made novel contributions by demonstrating that computational language models can produce features relevant for characterizing human language production to predict the degree of cognitive impairment in AD (i.e., the relation between surprisal and MMSE score). This motivates future work that could combine these models with automated feature-selection pipelines to identify those features that are most informative, similar to the approach employed by Fraser and colleagues. These models would jointly benefit from the inclusion of the new features that we have demonstrated are relevant—namely, lexical surprisal—and the data-driven feature selection process used by Fraser et al. to obtain peak classification performance.

Finally, in addition to reporting area under the ROC curve values and accuracies, we also examined precision, recall, and the impact of moving our model's decision

boundary to differentially weight true versus false positives. When increased recall was prioritized (the ratio of correctly predicted positives to the total number of actual positives), our model could, on average, identify approximately 87% of patients with AD. This highlights that the current approach could be tailored to function as a warning system that prioritizes sensitivity, which may be particularly useful for individuals who have a heightened risk of developing AD due to family history or other reasons.

Toward Future Speech-Based Models for Dementia Detection

There is a growing need for more accessible dementia screening systems, which do not require substantial time, travel, or financial resources for patients to access (Bayly et al. 2020; Bradford et al., 2009; Wilson et al., 2023). The collection and analysis of spoken language is one such promising method. Once speech samples have been collected, there are two analysis steps that have traditionally relied on time-consuming manual effort, namely, the transcription of spoken language to text and the computation of relevant features. While the data set in the present work was created using manual speech transcription, ongoing advances in automatic speech recognition suggest that this step could be partly or fully automated soon (e.g., Coto-Solano et al., 2021). For example, recent work (Liu et al., 2023) has used the TalkBank databases to develop and validate a largely automated transcription pipeline for converting raw audio input into speech transcripts in the CHAT format, which facilitates subsequent analysis. Others have compared automated and manual speech transcription methods for the extraction of linguistic features and found similar accuracies when using the resulting features to classify AD patients from healthy controls (Sadeghian et al., 2021; see Li et al., 2024, for related work). Importantly, all the linguistic features that were used in the current work were extracted and computed automatically from speech transcripts, with no manual annotation required. This includes the surprisal-based features developed in this work, which, our results suggest, should be included in future models that aim to detect AD in earlier disease stages and/or predict cognitive decline.

This is, as far as we know, the first study to consider the relationship between lexical surprisal and cognitive status in dementia. One reason may be that, until recently, surprisal was a difficult metric to compute automatically. Traditionally, when analyzing language production, word probability was computed using cloze probability norming, a method wherein a separate norming sample of participants are given an incomplete sentence and guess what word will appear next (e.g., *The man put on his* ____; Taylor, 1953). This approach requires substantial resources and time, as norming participants would need to guess each word in every transcript. Here, we circumvent this manual effort by taking advantage of advances in computational language modeling and the widespread availability of pretrained large language models. Using one such model and relying on its statistical "knowledge" of the English language, derived from training on millions of documents, we can automatically estimate the contextual probability of each word in a transcribed speech sample. We note that the current work is only an initial step toward using these models to improve speech-based screening tools and many open questions remain. For example, the current work employed the GPT-2 language model to derive the lexical predictions for computing surprisal. However, the field of large language models is moving rapidly, and already the suite of computational language models has greatly expanded and been refined with new training data (e.g., GPT-4, Llama). While most of these models have been trained on written language, they could be fine-tuned using corpora of transcribed, spoken language, which may improve their ability to detect deviations from expected speech patterns. This future research will be facilitated by the continued public availability of pretrained models.

Limitations and Future Directions

The present approach employed one computational language model, GPT-2, to estimate the incremental probability of individual words in the speech stream. Although this model was deliberately chosen due to various properties of its training, and correlations with real human language behavior, we do not wish to claim that GPT-2 is the best model to derive these probabilities for the purposes of predicting cognitive impairment. A natural line for future work will be to test linguistic features derived from different computational language models and modify their associated parameters (e.g., window size), to determine how these variables affect performance when predicting degree of cognitive impairment in dementia patients.

The current work is also limited due to our reliance on the precollected DementiaBank data set (Becker et al., 1994). This publicly available data set is the only one of its kind and is a uniquely valuable hypothesis testing ground for several reasons, including its large sample size (an order of magnitude larger than the sample size in many similar studies) and the diversity of MMSE scores among the AD patient sample. It does not, however, provide researchers with comprehensive patient descriptions that could better inform our understanding of how linguistic changes relate to disease progression. In the present

work, we relied exclusively on the original authors' classification of patients as "Probable" or "Possible" AD cases and operationalized their degree of impairment using performance on the MMSE (Folstein, Folstein, & McHugh, 1975). Although MMSE performance is expected to decrease with greater impairment in dementia patients, it is ultimately only a screening tool and should not be used as a singular diagnostic tool (Tombaugh & McIntyre, 1992), as it has been found to vary considerably in its accuracy and sensitivity (Arevalo-Rodriguez et al., 2015, 2021; see earlier discussion in Data Set section). Future work should thus examine speech patterns in more precisely defined subclasses of dementia patients and use more comprehensive and updated neuropsychological evaluations as measures of impairment. This may be particularly fruitful in longitudinal studies that also quantify alterations in cortical microstructure, which can occur before overt brain atrophy is observed (Illán-Gala et al., 2019; Spotorno et al., 2023). These noninvasive methods have been shown to distinguish mild cognitive impairment and AD from typical aging, with localized differences between patients and healthy controls found in brain regions known to support language function (Vogt et al., 2020). If combined with regular speech assessments and thorough neuropsychological evaluations, such studies could improve our understanding of how speech changes relate to the underlying disease progression.

Another limitation of the current work concerns our reliance on the picture description task data available in the public data set. This task places a stimulus in front of each participant and asks them to describe what is depicted. As such, perception and memory demands are intertwined, as participants need to perceive and identify the objects and entities that make up the depicted scene, retrieve their names from memory, and then produce structured propositions that describe those entities and the depicted relations between them. While picture description tasks have been shown to have predictive power for future cognitive impairment, recent findings also suggest that linguistic features derived from alternative speech tasks, such as asking participants to describe a time or place in their life that is meaningful to them, may provide better predictive performance (Ostrand & Gunstad, 2021). Related work, which has examined speech collected in autobiographical interviews (Levine et al., 2002), has also shown that patients with AD and mild cognitive impairment produce fewer episodic memory details in expository speech as compared to healthy age-matched controls (Simpson et al., 2023). This raises another interesting question for future work, regarding whether different linguistic features best predict impairment when calculated from utterances describing episodic versus semantic information.

Conclusions

The current work makes three contributions to our understanding of linguistic changes in AD, the relationship between these changes and degree of cognitive decline, and the advancement of spoken language-based tools for the detection of AD. First, we have demonstrated that both word-level and contextually based properties of language can accurately predict AD patients' MMSE scores, which are expected to decrease with greater degree of cognitive decline. Second, we have shown that these features can be used to accurately classify patients with probable or possible AD versus healthy control participants, with correct detection of impairment reaching approximately 87% based on linguistic features alone. Finally, we have provided new evidence that linguistic surprisal—a contextual measure of how coherent or unexpected an individual's word choices are given the prior context—explains a unique proportion of variance in MMSE scores in AD patients, beyond that captured by many previously studied properties of language. This result suggests that linguistic surprisal should be studied and incorporated into future research that examines linguistic changes in dementia along the clinical staging for individuals on the AD continuum (Jack et al., 2024), as well as other forms of dementia, and particularly in longitudinal studies that can further assess its usefulness in early detection of disease. Overall, our results demonstrate that more severe dementia from AD leads to the production of more frequent yet more surprising words, greater use of empty words, and a reduced use of nouns and definite articles. We also further demonstrate that the automatic computation of properties of spontaneous speech can be an effective and low burden way of detecting and monitoring cognitive decline in AD.

Data Availability Statement

The data used in this research are publicly available in the TalkBank corpus (http://www.talkbank.org).

Acknowledgments

This research was supported by National Institutes of Health Grants R01AG065432 and P01AG073090 to Rachel Ostrand. Graham Flick was additionally supported by a Postdoctoral Fellowship from the Natural Sciences and Engineering Research Council of Canada. The acquisition of the original DementiaBank data was supported by National Institutes of Health Grants AG005133 and AG003705 to the University of Pittsburgh.

References

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain, 136*(12), 3727–3737. https://doi.org/10.1093/brain/awt269

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., & Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and Language, 67*(3), 202–227. https://doi.org/10.1006/brln.1999.2055

Alzheimer's Association. (2023). 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 19*(4), 1598–1695. https://doi.org/10.1002/alz.13016

Ambrosini, E., Cid, M., de Isla, C. G., Salamanca, P., Borghese, N. A., Ferrante, S., Caielli, M., Milis, M., Loizou, C., Azzolino, D., Damanti, S., Bertagnoli, L., Cesari, M., & Moccia, S. (2019). Automatic speech analysis to early detect functional cognitive decline in elderly population. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 212–216). https://doi.org/10.1109/EMBC.2019.8856768

Ambrosini, E., Giangregorio, C., Lomurno, E., Moccia, S., Milis, M., Loizou, C., Azzolino, D., Cesari, M., Cid Gala, M., Galán de Isla, C., Gomez-Raja, J., Borghese, N. A., Matteucci, M., & Ferrante, S. (2024). Automatic spontaneous speech analysis for the detection of cognitive functional decline in older adults: Multilanguage cross-sectional study. *JMIR Aging, 7,* Article e50537. https://doi.org/10.2196/50537

Arevalo-Rodriguez, I., Smailagic, N. I., Figuls, M. R., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O. L., Bonfill Cosp, X., & Cullum, S. (2015). Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews, 2015*(3), Article CD010783. https://doi.org/10.1002/14651858.CD010783.pub2

Arevalo-Rodriguez, I., Smailagic, N. I., Roqué-Figuls, M., Ciapponi, A., Sanchez-Perez, E., Giannakou, A., Pedraza, O. L., Bonfill Cosp, X., & Cullum, S. (2021). Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews, 7*(7), Article CD010783. https://doi.org/10.1002/14651858.CD010783.pub3

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

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition, 125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Bayles, K. A., Tomoeda, C. K., & Trosset, M. W. (1992). Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. *Brain and Language, 42*(4), 454–472. https://doi.org/10.1016/0093-934X(92)90079-T

Bayly, M., Morgan, D., Froehlich Chow, A., Kosteniuk, J., & Elliot, V. (2020). Dementia-related education and support service availability, accessibility, and use in rural areas: Barriers and solutions. *Canadian Journal on Aging / La Revue Canadienne Du Vieillissement, 39*(4), 545–585. https://doi.org/10.1017/S0714980819000564

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

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,* Article 369. https://doi.org/10.3389/fnagi.2018.00369

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1st ed.). O'Reilly Media.

Bird, H., Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language, 73*(1), 17–49. https://doi.org/10.1006/brln.2000.2293

Black, C. M., Fillit, H., Xie, L., Hu, X., Kariburyo, M. F., Ambegaonkar, B. M., Baser, O., Yuce, H., & Khandker, R. K. (2017). Economic burden, mortality, and institutionalization in patients newly diagnosed with Alzheimer's disease. *Journal of Alzheimer's Disease, 61*(1), 185–193. https://doi.org/10.3233/JAD-170518

Blanken, G., Dittmann, J., Haas, J.-C., & Wallesch, C.-W. (1987). Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. *Cognition, 27*(3), 247–274. https://doi.org/10.1016/S0010-0277(87)80011-2

Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P., & Singh, H. (2009). Missed and delayed diagnosis of dementia in primary care: Prevalence and contributing factors. *Alzheimer Disease & Associated Disorders, 23*(4), 306–314. https://doi.org/10.1097/WAD.0b013e3181a6bebc

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE, 14*(1), Article e0207741. https://doi.org/10.1371/journal.pone.0207741

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology, 14*(1), 71–91. https://doi.org/10.1080/026870300401603

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). *API design for machine learning software: Experiences from the scikit-learn project*. arXiv. http://arxiv.org/abs/1309.0238

Burke, D. M., & Shafto, M. A. (2008). Language and aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 373–443). Psychology Press.

Carlomagno, S., Santoro, A., Menditti, A., Pandolfi, M., & Marini, A. (2005). Referential communication in Alzheimer's type dementia. *Cortex, 41*(4), 520–534. https://doi.org/10.1016/S0010-9452(08)70192-8

Caucheteux, C., Gramfort, A., & King, J. R. (2021). Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning, 139*, 1336–1348.

Chen, Y., Hartsuiker, R. J., & Pistono, A. (2025). A comparison of different connected-speech tasks for detecting mild cognitive

impairment using multivariate pattern analysis. *Aphasiology, 39*(4), 476–499. https://doi.org/10.1080/02687038.2024.2358556

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004). *Fisher English Training Speech Part 1 transcripts LDC2004T19*. Linguistic Data Consortium.

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2005). *Fisher English Training Part 2, transcripts LDC2005T19*. Linguistic Data Consortium.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Coto-Solano, R., Stanford, J. N., & Reddy, S. K. (2021). Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with DARLA. *Frontiers in Artificial Intelligence, 4,* Article 662097. https://doi.org/10.3389/frai.2021.662097

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language, 53*(1), 1–19. https://doi.org/10.1006/brln.1996.0033

Cuetos, F., Arango-Lasprilla, J. C., Uribe, C., Valencia, C., & Lopera, F. (2007). Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease. *Journal of the International Neuropsychological Society, 13*(3), 433–439. https://doi.org/10.1017/S1355617707070609

Dalton, S. G. H., Kim, H., Richardson, J. D., & Wright, H. H. (2020). A compendium of core lexicon checklists. *Seminars in Speech and Language, 41*(1), 45–60. https://doi.org/10.1055/s-0039-3400972

Dalton, S. G. H. & Richardson, J. D. (2015). Core-lexicon and main-concept production during picture-sequence description in adults without brain damage and adults with aphasia. *American Journal of Speech-Language Pathology, 24*(4), S923–S938. https://doi.org/10.1044/2015_ajslp-14-0161

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Bidirectional encoder representations from transformers*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

de Vugt, M. E., & Verhey, F. R. J. (2013). The impact of early dementia diagnosis and intervention on informal caregivers. *Progress in Neurobiology, 110,* 54–62. https://doi.org/10.1016/j.pneurobio.2013.04.005

Diaz-Asper, C., Chandler, C., Turner, R. S., Reynolds, B., & Elvevåg, B. (2021). Acceptability of collecting speech samples from the elderly via the telephone. *DIGITAL HEALTH, 7*. https://doi.org/10.1177/20552076211002103

Ehrlich, J. S., Obler, L. K., & Clark, L. (1997). Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type. *Journal of Communication Disorders, 30*(2), 79–99. https://doi.org/10.1016/0021-9924(95)00053-4

Feyereisen, P., Berrewaerts, J., & Hupet, M. (2007). Pragmatic skills in the early stages of Alzheimer's disease: An analysis by means of a referential communication task. *International Journal of Language & Communication Disorders, 42*(1), 1–17. https://doi.org/10.1080/13682820600624216

Filiou, R. P., Bier, N., Slegers, A., Houzé, B., Belchior, P., & Brambati, S. M. (2019). Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: A scoping review. *Aphasiology, 34*(6), 723–755. https://doi.org/10.1080/02687038.2019.1608502

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. https://doi.org/10.1016/0022-3956(75)90026-6

Forbes, K. E., Venneri, A., & Shanks, M. F. (2002). Distinct patterns of spontaneous speech deterioration: An early predictor of Alzheimer's disease. *Brain and Cognition, 48*(2–3), 356–361.

Forbes-McKay, K., Shanks, M. F., & Venneri, A. (2013). Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. *Acta Neuropsychiatrica, 25*(6), 320–327. https://doi.org/10.1017/neu.2013.16

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language, 140,* 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Fraser, K. C., Meltzer, J. A., Rudzicz, F., & Garrard, P. (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Godfrey, J., & Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Linguistic Data Consortium.

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Schain, M., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., . . . Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications, 15*(1), Article 2768. https://doi.org/10.1038/s41467-024-46631-y

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, S. A., Casto, C., Fanda, L., Doyle, W., . . . Hasson, U. (2020). Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv,* 2020–12. https://doi.org/10.1101/2020.12.02.403477

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *The Boston Diagnostic Aphasia Examination: BDAE-3 Long Form Kit–Third Edition*. Lippincott Williams & Wilkins.

Guinn, C. I., & Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. In *Proceedings of AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*.

Hier, D. B., Hagenlocker, K., & Shindler, A. G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and Language, 25*(1), 117–133. https://doi.org/10.1016/0093-934X(85)90124-5

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin, 7*(2), 172–177.

Illán-Gala, I., Montal, V., Borrego-Écija, S., Vilaplana, E., Pegueroles, J., Alcolea, D., Sánchez-Saudinós, B., Clarimón, J., Turón-Sans, J., Bargalló, N., González-Ortiz, S., Rosen, H. J., Gorno-Tempini, M. L., Miller, B. L., Lladó, A., Rojas-García, R., Blesa, R., Sánchez-Valle, R., Lleó, A., & Fortea, J. (2019). Cortical microstructure in the behavioural variant of frontotemporal dementia: Looking beyond atrophy. *Brain, 142*(4), 1121–1133. https://doi.org/10.1093/brain/awz031

Jack, C. R., Jr., Andrews, J. S., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., Ho, C., Jagust, W., McDade, E., Molinuevo, J. L., Okonkwo, O. C., Pani, L., Rafii, M. S., Scheltens, P., Siemers, E., Snyder, H. M., Sperling, R., Teunissen, C. E., & Carrillo, M. C. (2024). Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's & Dementia, 20*(8), 5143–5169. https://doi.org/10.1002/alz.13859

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014). Aided diagnosis of

dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 27–37). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3204

Kavé, G., & Dassa, A. (2018). Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology, 32*(1), 27–40. https://doi.org/10.1080/02687038.2017.1303441

Kavé, G., & Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 38*(9), 958–966. https://doi.org/10.1080/13803395.2016.1179266

Kavé, G., & Goral, M. (2018). Word retrieval in connected speech in Alzheimer's disease: A review with meta-analyses. *Aphasiology, 32*(1), 4–26. https://doi.org/10.1080/02687038.2017.1338663

Kemper, S., & Altmann, L. J. P. (2017). Dementia and language. In *Reference module in neuroscience and biobehavioral psychology*. Elsevier. https://doi.org/10.1016/b978-0-12-809324-5.01884-8

Kintz, S., Kim, H., & Wright, H. H. (2024). A preliminary investigation on core lexicon analysis in dementia of the Alzheimer's type. *International Journal of Language & Communication Disorders, 59*(4), 1336–1350. https://doi.org/10.1111/1460-6984.12999

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging, 17*(4), 677–689. https://doi.org/10.1037/0882-7974.17.4.677

Li, C., Xu, W., Cohen, T., & Pakhomov, S. (2024). Useful blunders: Can automated speech recognition errors improve downstream dementia classification? *Journal of Biomedical Informatics, 150,* Article 104598. https://doi.org/10.1016/j.jbi.2024.104598

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research, 66*(7), 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. *Proceedings of Interspeech, 2020,* 2172–2176. https://doi.org/10.21437/Interspeech.2020-2571

Luz, S., Haider, F., de la Fuente, S, Fromm, D., & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSSo challenge. *Proceedings of Interspeech 2021,* 3780–3784. https://doi.org/10.21437/Interspeech.2021-1220

MacWhinney, B. (2000). *The Childes Project: Tools for analyzing talk. Transcription format and programs*. Psychology Press.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

McDowell, I., Kristjansson, B., Hill, G. B., & Hébert, R. (1997). Community screening for dementia: The Mini Mental State Exam (MMSE) and Modified Mini-Mental State Exam (3MS) compared. *Journal of Clinical Epidemiology, 50*(4), 377–383. https://doi.org/10.1016/s0895-4356(97)00060-7

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*(7), 939–944. https://doi.org/10.1212/wnl.34.7.939

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., & Arana, J. M. (2014). Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders, 37*(5–6), 327–334. https://doi.org/10.1159/000356726

Misra, K. (2022). *minicons: Enabling flexible behavioral and representational analyses of transformer language models*. arXiv. http://arxiv.org/abs/2203.13112

Mitchell, A. J. (2017). The Mini-Mental State Examination (MMSE): Update on its diagnostic accuracy and clinical utility for cognitive disorders. In A. J. Larner (Ed.), *Cognitive screening instruments: A practical approach* (pp. 37–48). Springer.

Mittelman, M. S., Ferris, S. H., Shulman, E., Steinberg, G., & Levin, B. (1996). A family intervention to delay nursing home placement of patients with Alzheimer disease. A randomized controlled trial. *Journal of the American Medical Association, 276*(21), 1725–1731. https://doi.org/10.1001/jama.276.21.1725

Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology, 40*(9), 917–939. https://doi.org/10.1080/13803395.2018.1446513

Mueller, K. D., Koscik, R. L., Clark, L. R., Hermann, B. P., Johnson, S. C., & Turkstra, L. S. (2018). The latent structure and test–retest stability of connected language measures in the Wisconsin Registry for Alzheimer's Prevention (WRAP). *Archives of Clinical Neuropsychology, 33*(8), 993–1005. https://doi.org/10.1093/arclin/acx116

Mueller, K. D., Koscik, R. L., Hermann, B. P., Johnson, S. C., & Turkstra, L. S. (2018). Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience, 9,* Article 437. https://doi.org/10.3389/fnagi.2017.00437

Mueller, K. D., Koscik, R. L., Turkstra, L. S., Riedeman, S. K., LaRue, A., Clark, L. R., Hermann, B., Sager, M. A., & Johnson, S. C. (2016). Connected language in late middle-aged adults at risk for Alzheimer's disease. *Journal of Alzheimer's Disease, 54*(4), 1539–1550. https://doi.org/10.3233/JAD-160252

Murray, J., Schneider, J., Banerjee, S., & Mann, A. (1999). EUROCARE: A cross-national study of co-resident spouse carers for people with Alzheimer's disease: II—A qualitative analysis of the experience of caregiving. *International Journal of Geriatric Psychiatry, 14*(8), 662–667. https://doi.org/10.1002/(SICI)1099-1166(199908)14:8<662::AID-GPS993>3.0.CO;2-4

Murray, L. L. (2010). Distinguishing clinical depression from early Alzheimer's disease in elderly people: Can narrative analysis help? *Aphasiology, 24*(6–8), 928–939. https://doi.org/10.1080/02687030903422460

Nicholas, M., Obler, L. K., Albert, M. L., & Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech and Hearing Research, 28*(3), 405–410. https://doi.org/10.1044/jshr.2803.405

Orimaye, S. O., Wong, J. S.-M., & Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,* 78–87. https://doi.org/10.3115/v1/W14-3210

Ostrand, R., & Gunstad, J. (2021). Using automatic assessment of speech production to predict current and future cognitive function in older adults. *Journal of Geriatric Psychiatry and Neurology, 34*(5), 357–369. https://doi.org/10.1177/0891988720933358

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia, 9*(1), 63–75.e2. https://doi.org/10.1016/j.jalz.2012.11.007

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* [Technical report]. OpenAI.

Rasmussen, J., & Langerman, H. (2019). Alzheimer's disease—Why we need early diagnosis. *Degenerative Neurological and Neuromuscular Disease, 9,* 123–130. https://doi.org/10.2147/DNND.S228939

Sadeghian, R., Schaffer, J. D., & Zahorian, S. A. (2021). Towards an automatic speech-based diagnostic test for Alzheimer's disease. *Frontiers in Computer Science, 3,* Article 624594. https://doi.org/10.3389/fcomp.2021.624594

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Siemers, E., Holdridge, K. C., Sundell, K. L., & Liu-Seifert, H. (2016). Function and clinical meaningfulness of treatments for mild Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2*(1), 105–112. https://doi.org/10.1016/j.dadm.2016.02.006

Simpson, S., Eskandaripour, M., & Levine, B. (2023). Effects of healthy and neuropathological aging on autobiographical memory: A meta-analysis of studies using the autobiographical interview. *The Journals of Gerontology: Series B, 78*(10), 1617–1624. https://doi.org/10.1093/geronb/gbad077

Slegers, A., Filiou, R.-P., Montembeault, M., Brambati, S. M., & Migliaccio, R. (2018). Connected speech features from picture description in Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease, 65*(2), 519–542. https://doi.org/10.3233/JAD-170881

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Stark, B. C., Dalton, S. G., & Lanzi, A. M. (2025). Access to context-specific lexical-semantic information during discourse tasks differentiates speakers with latent aphasia, mild cognitive impairment, and cognitively healthy adults. *Frontiers in Human Neuroscience, 18,* Article 1500735. https://doi.org/10.3389/fnhum.2024.1500735

Spotorno, N., Strandberg, O., Vis, G., Stomrud, E., Nilsson, M., & Hansson, O. (2023). Measures of cortical microstructure are linked to amyloid pathology in Alzheimer's disease. *Brain, 146*(4), 1602–1614. https://doi.org/10.1093/brain/awac343

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., & Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005, 3,* 1569–1574. https://doi.org/10.1109/ICMA.2005.1626789

Tombaugh, T. N., & McIntyre, N. J. (1992). The Mini-Mental State Examination: A comprehensive review. *Journal of the American Geriatrics Society, 40*(9), 922–935. https://doi.org/10.1111/j.1532-5415.1992.tb01992.x

Vogt, N. M., Hunt, J. F., Adluru, N., Dean, D. C., III, Johnson, S. C., Asthana, S., Yu, J.-P. J., Alexander, A. L., & Bendlin, B. B. (2020). Cortical microstructural alterations in mild cognitive impairment and Alzheimer's disease dementia. *Cerebral Cortex, 30*(5), 2948–2960. https://doi.org/10.1093/cercor/bhz286

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., & Jack, C. R., Jr. (2007). 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. *Brain, 130*(Pt. 7), 1777–1786. https://doi.org/10.1093/brain/awm112

Wilson, N.-A., Peters, R., Lautenschlager, N. T., & Anstey, K. J. (2023). Testing times for dementia: A community survey identifying contemporary barriers to risk reduction and screening. *Alzheimer's Research & Therapy, 15*(1), Article 76. https://doi.org/10.1186/s13195-023-01219-4

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. M. (2020). *HuggingFace's transformers: State-of-the-art natural language processing*. arXiv. http://arxiv.org/abs/1910.03771