Supplementary Material for “Linguistic markers predict onset of Alzheimer’s Disease”

# Selection of Participants and Samples

## Exclusion Criteria

Application of the exclusion criteria involved excluding participants with a history of Parkinson’s disease, and excluding samples that could have been collected after a stroke. We identified participants with a history of Parkinson’s disease through searching for the keyword “Parkinson” in their list of medications. We identified participants with a history of stroke from the document titled “Survival file and follow-up for cardiovascular events”. If the value of the stroke variable in this document was 1, we used the earliest stroke age as the cut-off age. If the value of the stroke variable was 0, we used the censor age + 1 as the cut-off age. Samples collected before cut-off ages were marked for exclusion. The samples of participants who were not included in the document “Survival file and follow-up for cardiovascular events” were not marked for exclusion.

The exclusion criteria was applied to the entire test data, but only to the control group of the training data. The fact that the training data has weak labels does allow for some degree of noise in this data set. Furthermore, it is unlikely that samples collected after a stroke or onset of Parkinson’s disease symptoms would be included in the case data set in large numbers, as it was designed to include samples collected before the onset of cognitive impairment. Even if such samples were included in the case set of the training data, this would only provide better separation of the two classes, as we expect that samples from the preclinical period to be closer to the pathological speech that would be expected from the subjects/samples that satisfy the exclusion criteria. Furthermore, as the number of cases in the training set was significantly lower than the controls, which was partially alleviated by not applying the inclusion criteria to the case group of the training. Therefore, we did not apply the exclusion criteria to the case group of the training set. Out of 106 samples from cases in the training data, 13 of them were collected after a stroke date. Out of these 13 samples, 10 of them had a dementia rating of 0 (no cognitive impairment), and 3 had a dementia rating of 0.5 (MCI). We believe that an extensive neuropsychological examination involving 13 different tests, some of which directly assess language abilities, would not result in dementia ratings of 0 or 0.5 for subjects with detectable levels of brain damage in the language processing areas of the brain.

## Test data - Overview

The onset of AD was defined as the onset of cognitive impairment (MCI) in a subject who later received a diagnosis of AD. AD patients who developed MCI on or before age 85 (85) were defined as *cases*. We identified the participants who had a record of not having dementia in the FHS database, and inferred the oldest age they were recorded to be dementia free from the database. The *normal-aging* group was defined as the participants who were recorded to be dementia free on or after age 85 ( 85).

The control group was defined as the combination of the normal-aging group and AD patients whose onset of cognitive impairment was after 85 (>85) years old. According to this definition, all cases have already developed cognitive impairment due to AD at 85, and none of the controls have developed cognitive impairment due to AD at 85.

Age 85 was chosen as a threshold, because this threshold was the optimum age to provide the largest balanced test set from the FHS data that was available to us. As the age threshold increases, less participants qualify to be controls, and more participants qualify to be cases. Conversely, as the age threshold decreases, more participants qualify to be controls, and less participants qualify to be cases. In addition to providing the largest test set from FHS, age 85 has been widely used as a threshold to define the oldest-old in AD studies. Notably, later age of onset is associated less with genetic factors and more with environmental factors 29 30 31. The findings from real world data in Optum with >5000 patient records show that most individuals who are diagnosed with AD have some cognitive impairment by age 85.

The test data set included only one sample per participant, and they were matched to a control sample using age (+/- 2 years), gender, and education (college degree vs no college degree). As our purpose is to be able to classify cognitively normal subjects, we included only samples collected prior to any cognitive impairment onset. Since the question of whether someone will develop AD symptoms by 85 is not meaningful to ask for someone older than 85, samples that were collected after age 85 were also excluded from the test set. Finally, all samples collected after the dementia review were excluded from the test set.

### Selection of a balanced test set

After participants were selected according to the inclusion criteria described above, samples from these participants were further selected if they were collected before onset of cognitive impairment, before age 85 and before the participant’s dementia review. For some subjects, this resulted in multiple eligible samples for inclusion in the test set. In order to create a balanced test-set which includes only one sample per participant, we used bipartite matching algorithm for selection of the samples 32. Bipartite matching algorithm maximizes the number of selected participants while satisfying the constraints for age, education and gender matching.

#### Cognitive status at death

To determine whether a participant was demented at death, we used autopsy data and dementia review data. In the document titled “Neuropathology at autopsy, original and offspring cohorts”, the CERAD neuropathological criteria was used for any degree of Alzheimer type pathology. The participants who were recorded as definite Alzheimer's disease, probable Alzheimer's disease, and possible Alzheimer's disease, according to NPCERAD variable (values 1,2 and 3) were labeled as “AD through autopsy”.

In addition, we used demrv096 (cognitive status at time of death) and demrv097 (certainty of cognitive status at death) variables from the dementia review. If certainty of cognitive status at death was above moderate (3 = moderately; 4 = reasonably; 5 = highly) and cognitive status at time of death was ‘demented’ (mild dementia (1), greater than or equal to mild dementia (1.5) , moderate dementia (2greater than or equal to moderate dementia (2.5); severe dementia (3)); then we labeled the participant as “demented at death according to dementia review”

If a participant was labeled “AD through autopsy” or “demented at death according to dementia review” then they were labeled as “demented at death”.

#### Alzheimer’s Disease

The participants who had Alzheimer's disease by NINCDS-ADRDA criteria were identified with the demrv115 variable in their dementia review. If the value of the demrv115 variable was “yes”, or the participant was labeled as ‘AD through autopsy’, as described above, then they were labeled as AD participants.

#### Not-demented status

The participants who were not demented according to DSM-IV criteria were identified with the demrv112 variable in their dementia review. If the value of this variable was “yes”, and the participant was not labeled as ‘demented at death”, as described above, then they were labeled “not-demented” participants.

#### Oldest record of not-demented status

For participants labeled as not-demented, as described above, we need to determine the oldest age they were recorded to be not demented in FHS.

We first determined whether a participant was alive at the time of dementia review. For this purpose, we used the demrv096 (cognitive status at time of death) variable from the review, which has value “alive” for some participants. Alternatively, if the age of a participant at their latest NP exam is greater than their age during their dementia review, this implies that the participant was alive during their dementia review. For participants who were thus determined to be alive during their dementia review, the review age was used as the oldest record of not-demented status, as AD diagnoses were based on their dementia review data, and any later change in participants’ cognitive status cannot be clinically defined without dementia review data.

For participants who were not alive during their dementia review, if their age of their death was recorded in the document titled “Neuropathology at autopsy, original and offspring cohorts”, we used the age of death as the oldest record of not-demented status, as their dementia review panel had their data available until death. Otherwise, we used the latest of (1) oldest NP exam age, (2) age of last date documented to be cognitively intact (normal\_date variable in dementia review) (3) age of cognitive impairment onset (impairment\_date variable in dementia review) as then the oldest record of not-demented status.

#### Test data - Summary

If a participant was labeled as “not-demented” as described above, and the “oldest record of not-demented status” was >= 85; or the participants was labeled as AD as described above, and their age of cognitive impairment onset, (impairment\_date) variable in dementia review was > 85, then they were labeled as controls.

If a participant was labeled as AD, as described above, and their age of cognitive impairment onset (impairment\_date variable in dementia review) was <= 85, then they were labeled as cases.

### Date of cognitive impairment onset in dementia review vs onset of MCI

The impairment\_date variable in dementia review stands for the date of onset of any cognitive impairment. We were not provided any variables that would determine whether this date stands for the onset of MCI or another type of impairment in the dementia review. This decision may cause some minor noise in our *test set*, as described below. For a given participant,

* If this date stands for the onset of MCI but not for another type of cognitive impairment, then this is the actual date of onset of MCI.
* If this date stands for the onset of a cognitive impairment other than MCI, then this date is a lower bound on the onset of MCI. This is because if MCI onset was *before* the onset of another cognitive impairment, then the date of MCI onset would have to be in dementia review as the impairment\_date, according to the definition of this variable. But this would be a contradiction. Therefore, the actual onset of MCI is after the impairment\_date, if this date stands for any cognitive impairment other than MCI.
* Following (1) and (2), date of MCI onset is >= impairment\_date. Therefore, the impairment\_date variable is a tight lower bound for date of MCI onset.
* Since we define the age threshold for inclusion in the oldest-old AD group in terms of an inequality, such that age of onset of MCI is >85, it is possible to use the impairment\_date variable for the definition of this group, as

MCI onset >= impairment\_date > 85

implies

MCI onset > 85.

* Therefore, no noise is added to the oldest-old AD group by using the impairment\_date variable for onset of MCI.

For the AD participants who are in the case group, the impairment\_date is <= 85. It is possible that a few of the case participants may have had cognitive impairment not due to MCI before 85, but got diagnosed with AD later with onset of MCI after 85. These participants should have been in the oldest-old AD group. We have good reason to believe that such participants, if they exists, are very few in number. Because, for such patients, the time between the impairment\_date and the diagnosis of mild AD should be longer than the expected time to convert from MCI to AD, as we assumed that cognitive impairment onset was before MCI. However, the median time to convert from MCI to AD of the test cases is 2.86 years. This shows that using the impairment\_date to define the test cases did not add significant noise to this data set. Furthermore, inclusion of an oldest-old participant in the case group by mistake would only make the classification decision harder, not easier, as they are more similar to the participants in the control group.

## Weakly-labeled Training Data

A second, larger cohort was identified to serve as training data, with inclusion criteria not requiring a dementia review. Participants included in the test set were excluded from the training set. Participants who qualified to be controls according to the test set selection criteria, but were excluded from the test set due to age-gender-education matching were included in the training set as controls.

For subjects without a dementia review, the training data labels were derived using dementia ratings for each administration of the NP test battery, with the mri352 variable. This categorical variable has the following values: not demented (0), not demented or mild cognitive impairment (MCI) (0.25), MCI (0.5), mildly demented (1), mildly to moderately demented (1.5), moderately demented (2), moderately to severely demented (2.5), severely demented (3), other (5). We used these ratings to create weak labels for participants who did not have a dementia review as follows: participants who had at least one dementia rating ≥ 0.5 before age 85 were labeled as cases and participants who did not have any cognitive impairment rating (dementia rating > 0) before age 85 and had at least one NP exam on or after age 85 were labeled as controls (to ensure they were followed at least up to 85 years old). If a participant received a dementia rating ≥ 1 (demented), and later in time received a dementia rating <1 (non-demented), than that participant was excluded from the training data set, since we assumed cognitive decline due to aging does not reverse its course.

For the control group, we included only samples with dementia rating of 0 (not-demented). For the case group, we included samples collected before the onset of cognitive impairment, if impairment\_date variable was set in dementia review. Otherwise, we analyzed the series of the dementia ratings to discover samples collected before the onset of MCI for inclusion in the case data. The dementia rating of 0.25 is an ambiguous label that could stand for either MCI or not-demented in FHS. Since the number of samples for the case group was small, we resorted to disambiguating the samples with dementia rating of 0.25 for the case group as follows: if a patient had a rating of 0.25 with a dementia rating of 0.5 (MCI) between the 0.25 and demented ratings, then we disambiguated the 0.25 rating as not-demented, since two NP exams are expected to be 4.4 years apart, which is longer than the average time it takes for a subject to develop AD from MCI. Otherwise, if a patient had a dementia rating of 0.25 without any dementia rating of 0.5 (MCI) between the 0.25 and the demented ratings, then we disambiguated the 0.25 rating as MCI.

To summarize, we included samples with a dementia rating indicating not-demented (dating of 0 or rating of 0.25 disambiguated as not-demented), that were collected before the first sample with a dementia rating indicating MCI (0.5 or 0.25 disambiguated as MCI). In other words, samples with a dementia rating 0 that were collected before the first exam with a dementia rating = 0.25, and samples with a dementia rating < 0.5 collected before the first exam with a dementia rating = 0.5 were included for the case group.

# Linguistic Variables

## Verbosity, Lexical Richness and Repetitiveness

In psycholinguistics studies on language function in dementia and AD, verbosity and lexical richness stand out as strong predictors of cognitive aging. Both measures show a decline with aging-related cognitive impairment 333435. Metrics that are functions of the total number of words, the total number of unique words, and the number of utterances are commonly used to assess lexical richness and verbosity.

A closely related concept to lexical richness and verbosity is repetitiveness. Repetitive behaviors have been shown to be a clinical feature of AD and related dementias 36. For example, repetitive speech that involves repetitive questioning, repetitive stories/statements, repetitive themes have been reported in patients with dementia 363738). Repetitiveness, and lexical richness are intrinsically related, as repetition of words in a language sample decreases the total number of unique words in the sample, therefore decreases lexical richness. See Supplementary Table 6 for the list of variables used for assessment of verbosity, lexical diversity and repetitiveness, along with their descriptions.

## Agraphia

Our data consist of written language samples, which permits automatic analysis of the writing performance of the participants. Studies on agraphia in dementia and in AD participants have shown that patients made more writing errors compared to controls 39 40. The agraphia variables address two aspects of the writing style: the misspellings and the formality of the style. To address the formality of writing style, we analyzed uppercasing and use of punctuation. See Supplementary Table 6 for the list of variables used for assessment of participants’ writing, along with their descriptions.

## Semantic Content

The Cookie-theft picture has been one of the most widely used pictures for eliciting narrative language in studies on aphasia. As a result, the contents of this picture, namely the objects (e.g. the window, the sink, the boy), and the actions (e.g. falling, standing, washing) have all been subject to intense inquiry with respect to their use by dementia and AD patients 414243. As people may refer to the objects in the picture in a multitude of ways, for example, the boy in the picture can be referred to as ‘the boy’, ‘the son’, ‘the kid’, etc., the concept of ‘information content units’ (ICU) have been used in the literature to group multiple ways of referring to the same object in the picture. The ICU’s in Cookie-theft picture have been defined and used in literature to analyze which contents of the picture have been mentioned differentially by patients and healthy controls.

Our goal was to determine whether the ICU objects and actions were mentioned in the samples, and whether they were mentioned by using the actual word of the ICU, for example whether the participant referred to the boy in the image with the word “boy”, as opposed to using another word with a similar meaning, e.g. “kid”.

For that purpose, we used word embeddings obtained from Glove 44, which represent each word as a vector of numbers. Given two words, the similarity of the words are quantified by the cosine similarity between the vectors of the two words. The cosine similarity between two vectors is between -1, and 1, indicating how similar the words are. For example: if the words are identical, then the cosine similarity score is 1. The smaller the cosine similarity, the less similar the word pair.

For each ICU, we identified the most similar word (word with the maximum cosine similarity) in the sample and used its cosine similarity score as a variable. Accordingly, if an object ICU was mentioned in the sample using the actual word of the ICU, then the value of this variable is 1. However, if the value of this variable is not 1 but still a higher score, then the ICU was mentioned in the sample using a different word with the similar meaning. If the value of this variable is very low, then the ICU was not mentioned in the sample.

The information content units of the objects and the actions in the cookie-theft image used in this study are given in Supplementary Table 6.45 Although suggested in the literature, we did not use the word ‘being’ as an action, as it is a very underspecified concept. We did not use the action ‘daydreaming’, as it is a very infrequent word. Both of these actions were not amenable to the word embedding approach utilized in this study. We also did not use the action ‘sitting’ as there is no subject sitting in the picture.

## Idea and Propositional Density

The ICU’s in the cookie-theft picture have been used to quantify idea density, which was defined as the total number of ICUs divided by total number of words in a language sample 4647. Propositional *idea density* is a related but separate metric 484950, which quantifies both syntactic and semantic complexity through analysis of valency of verbs in a sentence. We used the CPIDR 51 software to compute propositionalidea density, which is the number of ideas divided by number of words. In addition, we used the total number of verbs and nouns divided by the total number of words as a proxy for idea density. See Supplementary Table 6 for the list of variables used for assessment of idea and propositional density, along with their descriptions.

## Word-frequencies and Language Modeling

The study of word-frequency in psycholinguistics dates back as far as 1950s 5253. Since then, word-frequencies have been used as a variable sensitive to and predictive of performance in fluency, competence, and difficulty of retrieval. As cognitive resources are diminished with aging, and more so in aging related neurodegenerative diseases, word-frequency affect has been observed in various aspects of language production of demented and AD patients 545556. Depending on the task, AD patients produced either higher frequency words, or lower frequency words as compared to healthy controls. Similar to frequencies of word usage, language-modeling research under NLP is tasked with modeling the distributions of word sequences, i.e. multi-word chunks. Language modeling related variables have been used successfully for automatic classification of patients with dementia 5. For example, language modeling can successfully model increased use of formulaic-language by AD patients 57, which would not be possible by modeling frequencies of words in isolation.

We computed word-frequencies using the NLTK tool for each word using the Brown corpus, and then used the median word frequency of words in a sample as a variable. We used the SRILM language modeling tool 58 to learn a language model from the DementiaBank corpus 59, as follows: We used modified Kneser-Ney smoothing for interpolation 60, and built an open vocabulary language model i.e., the model contained the unknown-word token as a regular word. The maximum order of the estimated language model was four. We then scored each sample using this learned model. We used the percentile 10, 50 and 90 of ngram probability scores as variables. In addition, we used the perplexity of the sample as a variable. Finally, we created a binary variable for each *n* in {2,3,4}, whether any of the multi-word chunks of length *n* occurred in the DementiaBank data set. The language modelling analyses were all performed on lemmatized data. See Supplementary Table 6 for the list of variables used for modeling frequencies of words and word sequences.

## Syntactic Complexity

Declines in structural complexity of utterances have been extensively investigated in people with Alzheimer’s disease and dementia. Demented participants were shown to have difficulties with both comprehension and production of syntactically complex sentences 616263. Although syntactic errors are relatively less frequent in the early stages of Alzheimer's disease, the utterances of the mildly demented adults were shown to be shorter and syntactically simpler than those produced by the non-demented adults 62. Frequently used metrics for quantifying syntactic complexity are the number of subordinate clauses, prepositions, main and secondary verbs, tense-aspect-modality markers, conjunctions, sentence length in clauses, lengths of paths on syntax trees, sub-tree patterns in syntax trees 6465666717.

For modeling syntactic complexity, we used the syntactic analysis presented in 68. Here, we are providing a summary of the method presented in this study:

### Summary of the method

The syntactic complexity analysis begins with obtaining syntactic parse trees of utterances. The syntactic parse tree can be created according to different grammar formalisms, and for this study, we used dependency grammar analysis, as presented in this section. Supplementary Figure 1 shows a dependency parse tree, which depicts dependency relations between word pairs with an outgoing arrow. In the relation between ‘this*’* and ‘paper’, the arrow originates from ‘paper’ and points to ‘this’. In dependency trees, an arrow originates from the head word, and points the dependent word, therefore ‘paper’ is the head, and ‘this’ is the dependent in this particular relation. In dependency trees, relations have types as indicated by the labels on the arrows, e.g. in Supplementary Figure 1, the relation between ‘this’ and ‘paper’ is a determiner (det) relation. The number of relations (edges) in a dependency tree equals the number of words, which in this example is 11.

We used the rates of relations per words, and the rates of pos-tags per utterances as variables. For example, there are two ‘det’ relations in Supplementary Figure 1 (arrows pointing to this and a), therefore the rate of ‘det’ relation per word is 2/11.

The out-degree of a node is the number of arrows originating from the node; for example, the out-degree of node method is 4 in Supplementary Figure 1. The depth of a dependency tree is the length of the longest path from its root to one of its leaves. We used standard deviation, median, percentile 10 and 90, skewness and kurtosis of the depth of the trees and the out-degrees of nodes as variables.

Finally, we considered relation-pairs that are consecutive in the dependency graph. Two arrows of a directed graph are consecutive if the arrow of the first one is at the nock of the second one. For example in Supplementary Figure 1, the relations ‘nsubj’ and ‘det’ are consecutive in this paper presents. Please note that the consecutive relation is not symmetric, e.g. ‘nsubj’ and ‘det’ are consecutive, but ‘det’ and ‘nsubj’ are not consecutive in this paper presents. Similarly, ‘root’ and ‘nsubj’ are consecutive in paper presents. In Supplementary Figure 1, some consecutive relation pairs occur more than once: for example the ‘dobj’ and ‘amod’ pair occurs twice: first in ‘independent method’, second in ‘syntactic complexity’. For each instance of a consecutive relation pair, we count the number of their occurrences, and normalize them by the total number utterances in the sample.

Supplementary Figure 1 shows the parse tree of a single sentence, however speech samples usually consist of multiple utterances. Consequently, we collected counts for each dependency-relation type, each consecutive relation pair, and each pos-tag across multiple utterances in a given sample.

### Elimination of correlated syntactic variables

We computed the syntactic variables on the DementiaBank corpus 59, which consists of 257 spoken samples of the cookie-theft task from 169 Alzheimer’s disease patients, and 242 samples from 99 controls. We analyzed this dataset to investigate the correlations between the syntactic variables. Since dependency-relation variables are more informative of the syntactic structure than the pos-tags, we eliminated all pos-tag variables that were correlated with a dependency-relation variable. The total number of pos-tag variables were 33, and 15 of them were eliminated due to a correlation with a dependency-relation variable. The correlated variables were determined by a Kendall tau correlation above 0.5 or below -0.5. 0.5 was chosen as a threshold, because information conveyed by pos-tags and dependency relations presumably overlap to a significant degree, and our aim was to eliminate redundancy as much as possible with a moderate threshold. Next, we eliminated consecutive-relation variables that were correlated with either a dependency-relation variable, or with a pos-tag variable that was not eliminated in the previous step. The correlated variables were determined by a Kendall tau correlation above 0.7 or below -0.7. We chose a higher threshold for elimination of consecutive-relation variables, because we believe that the information they convey is distinct from the information conveyed by the dependency-relation variables. The total number of consecutive-relation variables were 401, and 68 percent of them were eliminated with this method.

We computed Wilcoxon signed rank test for each variable between samples of the controls and patients in the DementiaBank dataset, and eliminated samples that had a p-value > 0.001. See Supplementary Table 6 for the complete list of syntactic variables used in our study.

# Non-linguistic variables

The hypertension variable was set to one if the participant’s data included a systolic blood pressure reading greater than 140, a diastolic blood pressure reading greater than 90, or use of antihypertensive medications. The diabetes indicator was set to one if the participant’s data included a non-fasting blood glucose reading greater than 200, a fasting blood glucose reading greater than 126, or use of anti-diabetic medication.

# Predictive Modeling Approach

The prediction experiments varied in the following dimensions: the variables used for prediction, and the machine learning methods used for assessing prediction performance. The variables used include the linguistic variables, the non-linguistic variables, and the combination of the two.

We demonstrate the predictive performance and generalization robustness of our method in two machine learning frameworks: the *cross-validation* method and the *hold-out* method. Both methods are based on the notion of a separation of samples used for learning the parameters of a predictive model (training set), and samples used for testing the prediction performance (test set). We chose to train a classifier for all combinations of predictive variable sets and validation methods, as it provides a simple interpretation of how predictive features are combined (linearly), a score that indicates the confidence of the prediction, and a continuous scale of how far from the decision boundary each sample is, allowing for additional insight.

## Cross-validation Method

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal size subsamples. Of the *k* subsamples, a single subsample is retained for testing, and the remaining *k*-1 subsamples are used for training. The cross-validation process is then repeated *k* times (the folds), with each of the *k* subsamples used exactly once as the test data. As a result, all samples are used for both training and validation, and each sample is used for validation exactly once. The *k* results from the folds can then be combined to produce an estimation of the entire data set. The cross-validation method has previously been used in automated linguistic analysis classifying AD patients 69 4 6. We set *k* = 20, and repeated the experiments 50 times. Variable selection was performed within each cross-validation fold, using only the data available within the fold for training. We used only the ground-truth labeled test data set in cross-validation experiments.

## Hold-out Validation Method

In this method, a separate training data set is used for variable selection and training machine learning models, while a fixed dataset is used for testing. We used the ground-truth labeled test set for testing, and the weakly-labeled training set only for training. The training data set was almost seven times larger than the test set. The classes were very heavily imbalanced; therefore, in order to alleviate this issue, the over sampling algorithm SMOTE 70 was performed on the training data before training, which resulted in 1034 samples for training. We used grid-search methods to learn the hyper parameters of the classifiers using nested cross-validation within the training set.

## Variable Selection

In total, 87 linguistic variables were computed. Since the test sets were balanced in terms of gender and education, these were not used as predictor variables in the classification experiments. Variable selection was performed by using a univariate test between the preclinical AD and the control group for each variable and eliminating variables that were not statistically significant (p >= 0.05). The univariate test that was used was the t-test in the cross-validation experiments and Wilcoxon signed rank test in the hold-out experiments. The use of different univariate tests for different experiment conditions was justified due to difference in data size, and the noise in the weak labels. The syntactic variables were further eliminated using an external data set as explained in the Supplementary material Section 3.6.2.

Variable selection was performed using the weakly labeled training data for the hold-out method and within the training data of each fold for the cross-validation method. This simple variable selection method was chosen because (1) it sufficiently decreased the number variables with respect to the number of samples, and (2) in the case of cross-validation experiments, it prevented leakage of information from the test set into the training, which is possible with using sophisticated variable selection methods.

We upsampled the training data to balance the number of samples in each class, which provided us with 1034 samples including the original and the artificial samples.  Correction for multiple comparisons over the actual 1034 training samples showed that all features that were statistically significant without correction were still statistically significant after the correction.

Supplementary Table 6 shows p-values and t-test scores of all variables used in the study, along with the means for the controls and the cases in the ground-truth labeled dataset. Tables 4 and 5, on the other hand, show the weights assigned to the variables by the best performing classifiers, which in turn were used for classification of test set. We performed a step-wise classification analysis by ranking the features with respect to the weights assigned to the features (shown in Supplementary Table 4). Supplementary Figure 4 shows the results of this analysis, and it shows that the highest AUC score was obtained with using the highest ranked 10 features.

Finally, Supplementary Table 3 shows variables that were statistically significant on both the training and the test sets at the same time, which shows that onset of AD was associated with telegraphic speech, repetitiveness, and misspellings.

## Modeling

We experimented with Naïve Bayes, Logistic Regression and Linear SVM classifiers. The hyperparameters of the classifiers were learned from the data using cross validation. The variable selection was performed first, followed by imputation of the missing variables. The mean values for imputation of missing non-linguistic variables were computed using the training data for both the training data and the test data. The mean for cases and controls were used for imputing the values of the training data. The overall mean values were used for imputing the missing values of the test data. The unobserved syntactic variables in the test set were filled with zeros.

Hyperparameter search was performed using cross-validation on the training data in the hold-out experiments, in that the best combination of hyperparameters were found using grid search over folds of the training data. Similarly, the hyperparameters search in the CV experiments was performed using *nested* cross-validation. In *nested* cross-validation, the best combination of hyperparameters were found using grid search the training data of each CV fold separately, and it was repeated for each fold of the CV experiment.

# Separation of the Test and Training Data Sets

The results of the model using only linguistic variables in the hold-out framework are shown in Supplementary Figure 2. Panels A and B show the classification of the samples in the training set. The blue dots represent the control participants, and the red dots represent the cases (preclinical AD participants). Participants whose profile fell below the horizontal line were classified as controls, and participants above the line were classified as cases. Further distance from the horizontal line indicates increased confidence in the classification of the participant. Panel C shows the distance to the decision hyperplane for all samples in the test set.

# The effects of education and sex on classification performance

To examine the effects of education and sex on performance of the model using linguistic variables and the hold-out method, AUC scores were computed for participants with college degree vs participants without a college degree, and for females vs males. The participants with a college degree were harder to classify than participants without a college degree (AUC of 0.70 for college-degree vs 0.76 for no-college degree). Similarly, females were both more accurately and more confidently classified than males, and the difference is substantial (AUC of 0.83 for females vs 0.64 for males). See Supplementary Figure 3 for the ROC curves of the entire test set and the ROC curves for education and gender grouped subjects in the test set.

# Classification Probability Estimate

In order to estimate the statistical significance of the AUC scores obtained with our prediction models, we followed the approach by 71, who suggest computing Mann-Whitney statistics against the distribution of all possible AUC values for the given numbers of cases and controls as a null hypothesis. In our case, given the relatively large sample set and the balance of case and control samples, we adapted the method by computing explicitly the null distribution, approximating it as Gaussian to compute its standard deviation, and then applying z-statistics to estimate the p-value of the given AUC against the null distribution.

*To estimate the probability of the null hypothesis, we computed the distribution of outcomes for random binary splits of two classes with n elements each. The probability of finding i elements of one class in a split is:*

**

**

*Where  is the binomial coefficient. From this we compute the probability of obtaining by chance a classification accuracy a as:*

**

*Where is the floor function, and Z the set of integer numbers. To estimate the gain over chance we compute a z-value as:*

**

*Where  and *

We then use standard z-statistics to compute the corresponding p-values.

# Plotting the ROC curves

The ROC curve illustrates the performance of a binary classifier system as its discrimination threshold is varied. The plot was created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. In classification tasks, each sample may be associated with a prediction along with a score for the prediction.  These scores are obtained from the probability estimates of the classes, or confidence values, such as distance to the separating hyperplane, depending on the particular classifier that computed the predictions and the scores. The scores typically have many tied values across samples. In the plot in our manuscript, the sorted scores with the distinct values were used as thresholds at which to compute TPR and FPR values. We used the scikit-learn implementation of the roc\_curve function to compute the thresholds, the TPR and FPR values at these thresholds.

# The Factors from the NMF Analysis

We computed the loading of the factors on the first component that was obtained from the non-negative matrix factorization analysis. The loading of the two groups of variables, the linguistic and the neuropsychological respectively, are shown in Supplementary Figure 5. Supplementary Figure 5 shows that the first component of the factorization analysis combines information from both sets of variables, and takes into account the correlational structure among these variables.

# Initialization and the hyperparameters of the classifiers

## Initialization parameters

"Naive\_Bayes": GaussianNB(),

**Logistic Regression Models:**

"LR1": SGDClassifier(random\_state=1952,loss='log', average = 3,

penalty='l1',

alpha=1e-3,

class\_weight='balanced')

"LR2": CDClassifier(penalty="l1",

loss="log",

multiclass=False,

max\_iter=200,

C=1,

tol=1e-3)

**SVM models:**

"SVM1 ": SGDClassifier(random\_state=1952,

class\_weight='balanced',

average = 3,

alpha=1e-3,

penalty='elasticnet')

"SVM2": LinearSVC(penalty='l2', C=1,# probability=True,

class\_weight='balanced')

## Hyperparameter search parameters

Naïve Bayes: None

**Logistic Regression Models:**

"LR1": {"alpha": np.logspace(-5, 2, 6)},

"LR2": {"alpha": np.logspace(-5, 2, 7)},

**SVM models:**

"SVM1": {"alpha": np.logspace(-5, 2, 6),

"l1\_ratio": 10\*\*np.array([-2, -1, -.5, -.25,

-.12, -.06, -.01])}

"SVM2": {"C": [0.1, 0.5, 1, 2],

"loss":['hinge', 'squared\_hinge']}

A picture containing food

Description automatically generated

**Supplementary Figure 1**: Example dependency parse tree

A screen shot of a computer

Description automatically generated

***Supplementary Figure 2****:* Separation of the classes in the weakly-labeled training set, and the ground-truth labeled test set. Blue dots represent samples from controls; red dots represent samples from cases. The x-axes represents the age of the participants when the CTT was administered. The y-axes are the distance from the separating hyperplane that classifies subjects; this distance indicates confidence in the classification. Dots above the line are classified as cases; dots below are classified as controls. Panels A and B show the classification of controls and cases, respectively, from the weakly-labeled training set. Panel C shows the performance of the same model on the held-out test set.

A screenshot of a map

Description automatically generatedA close up of a map

Description automatically generatedA close up of a map

Description automatically generated

Supplementary **Figure 3**: Test set ROC curves. Left panel: AUC result for the entire test set. Middle panel: AUC results for college (red) and non-college educated participants (green). Right panel: AUC results for female (red) and male participants (green).

A screenshot of a cell phone

Description automatically generated

Supplementary **Figure 4**: Step-wise classification of the test set, using the variables that were ranked according to their weights. The ranking of the variables, along with their weights are shown in Supplementary Table 4. The plot shows that highest AUC was obtained with using the highest ranked 10 variables.

![A close up of a logo

Description automatically generated]()

*Supplementary* ***Figure 5***: The loading of the factors on the first component from the non-negative matrix factorization analysis, which shows the respective contribution of two groups of variables in the computation of the plot in Figure 5 in the manuscript.

# Neuropsychological Examination

Supplementary Table 1: Neuropsychological Variables

|  |  |
| --- | --- |
| **Variable Description** | Range |
| Correct score with phonemic cue from 10 items of Boston Naming test | 0-4 |
| Correct score with phonemic cue from 30 items of Boston Naming test | 0-15 |
| Correct score with semantic cue from 10 items of Boston Naming test | 0-3 |
| Correct score with semantic cue from 30 items of Boston Naming test | 0-6 |
| Correct score without cue from 10 items of Boston Naming test | 0-10 |
| Correct score without cue from 30 items of Boston Naming test | 0-30 |
| Paired Associate Learning - Immediate | 0-21 |
| Easy Score from Paired Associate Learning - Delayed Recall | 0-6 |
| Easy Score from Paired Associate Learning - Immediate Recall | 0-18 |
| Paired Associate Learning - Delayed | 0-10 |
| Hard Score from Paired Associate Learning - Delayed Recall | 0-4 |
| Hard Score from Paired Associate Learning - Immediate Recall | 0-12 |
| Paired Associate Learning - Recognition | 0-10 |
| Highest string from Digits Backward Span | 0-8 |
| Highest string from Digits Forward Span | 0-9 |
| Logical Memory - Delayed recall | 0-23 |
| Logical Memory - Immediate recall | 0-23 |
| Logical Memory - Recognition | 0-11 |
| Raw Score from Wide Range Achievement Test - Reading | 15-57 |
| The total words (not includes the animal part) from the Verbal Fluency Test | 0-95 |
| The total words (only the animal part) from the Verbal Fluency Test | 0-45 |
| Time in minutes to finish Trail A | 0-7 |
| Time in minutes to finish Trail B | 0-10 |
| Total raw score from Similarities Test | 0-26 |
| Total score from Hooper Visual Organization Test | 0-30 |
| Visual Reproductions - Delayed Recall | 0-14 |
| Visual Reproductions - Immediate Recall | 0-14 |
| Visual Reproductions - Recognition | 0-4 |
| Finger taps right hand | 0-77.6 |
| Finger taps left hand | 0-74 |
| Information | 0-29 |
| Block design | 0-32 |

Supplementary ***Table 2:*** Classification results obtained by the classifiers, as defined in Section 10 of this document.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Models** | **Metrics** | **NAÏVE BAYES** | **LR1** | **LR2** | **SVM1** | **SVM2** |
| **Hold-out** | Linguistic | Accuracy | 0.59 | **0.7** | 0.71 | 0.68 | 0.65 |
| **Hold-out** | Linguistic | Sensitivity | 0.65 | **0.62** | 0.65 | 0.6 | 0.62 |
| **Hold-out** | Linguistic | Positive predictive value | 0.58 | **0.74** | 0.74 | 0.71 | 0.66 |
| **Hold-out** | Linguistic | AUC | 0.65 | **0.74** | 0.72 | 0.74 | 0.72 |
| **Hold-out** | Non-linguistic | Accuracy | 0.51 | 0.56 | **0.59** | 0.55 | 0.57 |
| **Hold-out** | Non-linguistic | Sensitivity | 0.38 | 0.38 | **0.48** | 0.35 | 0.42 |
| **Hold-out** | Non-linguistic | Positive predictive value | 0.52 | 0.6 | **0.61** | 0.58 | 0.61 |
| **Hold-out** | Non-linguistic | AUC | 0.52 | 0.6 | **0.6** | 0.6 | 0.6 |
| **Hold-out** | Combined | Accuracy | **0.69** | 0.6 | 0.57 | 0.57 | 0.57 |
| **Hold-out** | Combined | Sensitivity | **0.62** | 0.38 | 0.42 | 0.35 | 0.35 |
| **Hold-out** | Combined | Positive predictive value | **0.71** | 0.68 | 0.61 | 0.64 | 0.64 |
| **Hold-out** | Combined | AUC | **0.67** | 0.64 | 0.63 | 0.63 | 0.65 |
| **CV experiments** | Linguistic | AUC | 0.67 | 0.67 | **0.73** | 0.66 | 0.67 |
| **CV experiments** | Linguistic | Accuracy | 0.63 | 0.62 | **0.65** | 0.62 | 0.61 |
| **CV experiments** | Linguistic | Sensitivity | 0.35 | 0.65 | **0.67** | 0.65 | 0.65 |
| **CV experiments** | Linguistic | Positive predictive value | 0.79 | 0.62 | **0.64** | 0.61 | 0.6 |
| **CV experiments** | Non-linguistic | AUC | **0.64** | 0.54 | 0.58 | 0.55 | 0.56 |
| **CV experiments** | Non-linguistic | Accuracy | **0.6** | 0.53 | 0.55 | 0.54 | 0.53 |
| **CV experiments** | Non-linguistic | Sensitivity | **0.44** | 0.54 | 0.51 | 0.55 | 0.54 |
| **CV experiments** | Non-linguistic | Positive predictive value | **0.64** | 0.53 | 0.56 | 0.54 | 0.53 |
| **CV experiments** | Combined | AUC | **0.72** | 0.65 | 0.68 | 0.65 | 0.68 |
| **CV experiments** | Combined | Accuracy | **0.67** | 0.63 | 0.64 | 0.63 | 0.63 |
| **CV experiments** | Combined | Sensitivity | **0.44** | 0.65 | 0.63 | 0.65 | 0.62 |
| **CV experiments** | Combined | Positive predictive value | **0.81** | 0.63 | 0.64 | 0.63 | 0.63 |

Supplementary **Table 3**: Variables that were statistically significant with the Wilcoxon rank sum test on the training set (p-value < 0.05), and had a p-value < 0.05 using a t-test on the test set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Direction** | **Interpretation** | **Description** | **T-test pval** | **Wilcoxon**  **pval** |
| **0** | Controls have/show more | Similarities Test | Total raw score | 0.0104 | 0.0002 |
| **1** | Controls have/show more | Paired Associate Learning | Total Score | 0.012 | 0.0177 |
| **2** | Controls have/show more | Logical Memory | Immediate recall | 0.0122 | 0.0016 |
| **3** | Cases have/show more | Telegraphic speech | Number of prepositions normalized by number of words | 0.0124 | 0.0459 |
| **4** | Controls have/show more | Logical Memory | Delayed recall | 0.0173 | 0.0001 |
| **5** | Cases have/show more | Repetitiveness | Part-of-speech tags repeated one word apart | 0.0203 | 0.0 |
| **6** | Controls have/show more | Non-telegraphic speech | Number of dependency label sequence (aux, root) normalized by total number of sentences | 0.0273 | 0.0013 |
| **7** | Controls have/show more | Non-telegraphic speech | Number of dependency label sequence (nsubj, root) normalized by total number of sentences | 0.0346 | 0.0069 |
| **8** | Controls have/show more | Non-telegraphic speech | Number of determiners normalized by number of words | 0.0404 | 0.0166 |
| **9** | Cases have/show more | Agraphia | Misspelling exists | 0.0442 | 0.0001 |

Supplementary ***Table 4:*** The weights assigned to the linguistic variables by the best performing classifier, ranked by absolute value of their weights.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Group** | **Subgroup** | **Description** | **Weight** |
| **0** | Linguistic | Writing | Misspelling exists | 27.16 |
| **1** | Linguistic | Syntactic | Number of dependency label aux normalized by total nubmer of words | 26.42 |
| **2** | Linguistic | Syntactic | Number of dependency label sequence (aux, root) normalized by total number of sentences | -24.78 |
| **3** | Linguistic | Writing | Uppercased letter at the beginning of a line, normalized by total number of lines | -20.15 |
| **4** | Linguistic | Syntactic | Number of dependency label nmod normalized by total nubmer of words | 16.61 |
| **5** | Linguistic | Language modeling | Percentile 50 of 1gram probabilities | -16.26 |
| **6** | Linguistic | Language modeling | Percentile 10 of 1gram probabilities | -16.15 |
| **7** | Linguistic | Verbosity - Diversity | Median number of characters in words | 15.38 |
| **8** | Linguistic | Word frequency effect | Median word frequencies | -15.23 |
| **9** | Linguistic | Semantic | Max distance to ICU "washing" | 13.56 |
| **10** | Linguistic | Syntactic | Number of dependency label det normalized by total nubmer of words | 11.11 |
| **11** | Linguistic | Repetitiveness | Part-of-speech tags repeated one word apart | 11.02 |
| **12** | Linguistic | Language modeling | 4gram observed in DementiaBank dataset | -9.45 |
| **13** | Linguistic | Writing | Number of misspelled words normalized by total number of words | -9.36 |
| **14** | Linguistic | Syntactic | Number of determiners normalized by number of words | 8.88 |
| **15** | Linguistic | Syntactic | Number of dependency label sequence (det, nsubj) normalized by total number of sentences | -8.49 |
| **16** | Linguistic | Syntactic | Number of dependency label sequence (nsubj, root) normalized by total number of sentences | 6.96 |
| **17** | Linguistic | Syntactic | Number of dependency label sequence (det, nmod) normalized by total number of sentences | -5.68 |
| **18** | Linguistic | Repetitiveness | Part-of-speech tags repeated zero words apart | -5.15 |
| **19** | Linguistic | Syntactic | Number of nouns normalized by number of words | 4.53 |
| **20** | Linguistic | Syntactic | Part-of-speech tag counts of VBZ normalized by sentences | -3.73 |
| **21** | Linguistic | Syntactic | Number of prepositions normalized by number of words | 3.13 |
| **22** | Linguistic | Language modeling | Percentile 90 of 1gram probabilities | 2.88 |
| **23** | Linguistic | Language modeling | Perplexity of the entire sample | -0.29 |

Supplementary ***Table 5:*** The weights assigned to the non-linguistic variables by the best performing classifier, ranked by absolute value of their weights.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Group** | **Subgroup** | **Description** | **Weight** |
| **0** | Demographic | Age | Age | -2.93 |
| **1** | Demographic | College | College | 1.54 |
| **2** | Medical | Diabetes | Diabetes | -1.48 |
| **3** | Neuropsychological | Paired Associate Learning | Total Score | -1.17 |
| **4** | Neuropsychological | Trail A | Time in minutes to finish, Maximum is 7 minutes | -0.8 |
| **5** | Neuropsychological | Wide Range Achievement Test | Raw Score from reading | -0.76 |
| **6** | Neuropsychological | Visual Reproductions | Immediate Recall | -0.68 |
| **7** | Neuropsychological | Verbal Fluency Test | The total words (not includes the animal part) | -0.53 |
| **8** | Neuropsychological | Logical Memory | Delayed recall | -0.43 |
| **9** | Medical | Hypertension | Hypertension | -0.32 |
| **10** | Neuropsychological | Paired Associate Learning | Easy Score | -0.27 |
| **11** | Neuropsychological | Digits Backward Span | Highest string | -0.26 |
| **12** | Neuropsychological | Boston Naming test | Correct score without cue from 30 items | -0.24 |
| **13** | Neuropsychological | Logical Memory | Recognition | 0.19 |
| **14** | Neuropsychological | The Wechsler Adult Intelligence Scale | Total correct answers | -0.15 |
| **15** | Neuropsychological | Similarities Test | Total raw score | -0.14 |
| **16** | Neuropsychological | Visual Reproductions | Recognition | 0.12 |
| **17** | Neuropsychological | Logical Memory | Immediate recall | -0.11 |
| **18** | Neuropsychological | Visual Reproductions | Delayed Recall | 0.0 |

Supplementary ***Table 6:*** All variables used in the study. The last two columns show the number of samples that were not used in computing the t-test scores and p-values. It could not be computed if the values were missing for either of both classes in the ground-truth labeled dataset.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group** | **Subgroup** | **Description** | **Mean - Cases** | **Mean - Controls** | **T-test stat** | **T-test p-val** | **N - Cases** | **N - Controls** |
| **0** | Demographic | Age | Age | 78.78 | 78.92 | -0.13 | 0.9 | 40.0 | 40.0 |
| **1** | Demographic | College | College | 0.42 | 0.42 | 0.0 | 1.0 | 40.0 | 40.0 |
| **2** | Demographic | Gender | Gender | -0.1 | -0.1 | 0.0 | 1.0 | 40.0 | 40.0 |
| **3** | Genetic | Apoe | Apoe | 32.0 | 31.45 | 0.61 | 0.55 | 40.0 | 40.0 |
| **4** | Medical | Diabetes | Diabetes | 0.42 | 0.4 | 0.22 | 0.82 | 40.0 | 40.0 |
| **5** | Medical | Hypertension | Hypertension | 0.95 | 0.9 | 0.84 | 0.4 | 40.0 | 40.0 |
| **6** | Neuropsychological | Block Design | Total Score | 19.26 | 21.36 | -1.62 | 0.11 | 31.0 | 28.0 |
| **7** | Neuropsychological | Boston Naming Test | Correct score with phonemic cue from 10 items | 0.33 | 0.3 | 0.2 | 0.84 | 21.0 | 20.0 |
| **8** | Neuropsychological | Boston Naming Test | Correct score with phonemic cue from 30 items | 1.9 | 1.9 | 0.01 | 0.99 | 21.0 | 20.0 |
| **9** | Neuropsychological | Boston Naming Test | Correct score with semantic cue from 10 items | 0.0 | 0.0 | N/A | N/A | 21.0 | 20.0 |
| **10** | Neuropsychological | Boston Naming Test | Correct score with semantic cue from 30 items | 0.52 | 0.45 | 0.37 | 0.72 | 21.0 | 20.0 |
| **11** | Neuropsychological | Boston Naming Test | Correct score without cue from 10 items | 9.25 | 9.5 | -1.43 | 0.16 | 40.0 | 40.0 |
| **12** | Neuropsychological | Boston Naming Test | Correct score without cue from 30 items | 24.67 | 25.9 | -1.08 | 0.28 | 21.0 | 20.0 |
| **13** | Neuropsychological | Digits Backward Span | Highest string | 4.55 | 4.67 | -0.54 | 0.59 | 40.0 | 40.0 |
| **14** | Neuropsychological | Digits Forward Span | Highest string | 6.2 | 6.35 | -0.53 | 0.6 | 40.0 | 40.0 |
| **15** | Neuropsychological | Finger Tapping | Mean of five trails from left hand | 35.39 | 34.53 | 0.37 | 0.71 | 18.0 | 15.0 |
| **16** | Neuropsychological | Finger Tapping | Mean of five trials from right hand | 36.72 | 36.73 | 0.0 | 1.0 | 18.0 | 15.0 |
| **17** | Neuropsychological | Hooper Visual Organization Test | Total score | 19.7 | 20.6 | -0.61 | 0.54 | 20.0 | 20.0 |
| **18** | Neuropsychological | Logical Memory | Delayed recall | 7.78 | 9.6 | -2.43 | 0.02 | 40.0 | 40.0 |
| **19** | Neuropsychological | Logical Memory | Immediate recall | 8.65 | 10.62 | -2.57 | 0.01 | 40.0 | 40.0 |
| **20** | Neuropsychological | Logical Memory | Recognition | 9.19 | 9.6 | -0.87 | 0.39 | 21.0 | 20.0 |
| **21** | Neuropsychological | Paired Associate Learning | Easy Score | 16.18 | 16.77 | -1.79 | 0.08 | 40.0 | 40.0 |
| **22** | Neuropsychological | Paired Associate Learning | Easy Score Delayed Recall | 5.81 | 6.0 | -2.12 | 0.04 | 21.0 | 20.0 |
| **23** | Neuropsychological | Paired Associate Learning | Hard Score | 3.58 | 5.12 | -2.5 | 0.01 | 40.0 | 40.0 |
| **24** | Neuropsychological | Paired Associate Learning | Hard Score from Delayed Recall | 1.48 | 2.2 | -1.89 | 0.07 | 21.0 | 20.0 |
| **25** | Neuropsychological | Paired Associate Learning | Total Score | 11.9 | 13.72 | -2.57 | 0.01 | 40.0 | 40.0 |
| **26** | Neuropsychological | Paired Associate Learning | Total Score from Delayed Recall | 7.29 | 8.2 | -2.26 | 0.03 | 21.0 | 20.0 |
| **27** | Neuropsychological | Paired Associate Learning | Total Score from Recognition | N/A | 10.0 | N/A | N/A | 0.0 | 2.0 |
| **28** | Neuropsychological | Similarities Test | Total raw score | 12.57 | 15.03 | -2.63 | 0.01 | 40.0 | 40.0 |
| **29** | Neuropsychological | The Wechsler Adult Intelligence Scale | Total correct answers | 18.09 | 20.1 | -1.67 | 0.1 | 32.0 | 29.0 |
| **30** | Neuropsychological | Trail A | Time in minutes to finish, Maximum is 7 minutes | 0.9 | 0.77 | 1.46 | 0.15 | 20.0 | 20.0 |
| **31** | Neuropsychological | Trail B | Time in minutes to finish, Maximum is 10 minutes | 3.28 | 2.05 | 2.39 | 0.02 | 20.0 | 20.0 |
| **32** | Neuropsychological | Verbal Fluency Test | The total words (not includes the animal part) | 33.15 | 32.48 | 0.26 | 0.8 | 40.0 | 40.0 |
| **33** | Neuropsychological | Verbal Fluency Test | The total words from the animal part | N/A | 18.5 | N/A | N/A | 0.0 | 2.0 |
| **34** | Neuropsychological | Visual Reproductions | Delayed Recall | 4.75 | 5.78 | -1.63 | 0.11 | 40.0 | 40.0 |
| **35** | Neuropsychological | Visual Reproductions | Immediate Recall | 6.2 | 6.72 | -0.77 | 0.44 | 40.0 | 40.0 |
| **36** | Neuropsychological | Visual Reproductions | Recognition | 1.92 | 2.25 | -1.41 | 0.16 | 39.0 | 40.0 |
| **37** | Neuropsychological | Wide Range Achievement Test | Raw Score from reading | 47.19 | 49.65 | -1.5 | 0.14 | 21.0 | 20.0 |
| **38** | Linguistic | Idea Density | Number of propositional ideas | 15.72 | 14.72 | 0.56 | 0.58 | 40.0 | 40.0 |
| **39** | Linguistic | Idea Density | Propositional idea density | 0.41 | 0.4 | 0.56 | 0.58 | 40.0 | 40.0 |
| **40** | Linguistic | Language Modeling | 2gram observed in DementiaBank dataset | 13.05 | 12.82 | 0.16 | 0.87 | 40.0 | 40.0 |
| **41** | Linguistic | Language Modeling | 3gram observed in DementiaBank dataset | 7.75 | 7.8 | -0.05 | 0.96 | 40.0 | 40.0 |
| **42** | Linguistic | Language Modeling | 4gram observed in DementiaBank dataset | 9.38 | 10.35 | -0.54 | 0.59 | 40.0 | 40.0 |
| **43** | Linguistic | Language Modeling | Percentile 10 of 1gram probabilities | 0.0 | 0.0 | -1.7 | 0.09 | 40.0 | 40.0 |
| **44** | Linguistic | Language Modeling | Percentile 10 of 2gram probabilities | 0.0 | 0.01 | -0.76 | 0.45 | 40.0 | 40.0 |
| **45** | Linguistic | Language Modeling | Percentile 10 of 3gram probabilities | 0.07 | 0.05 | 1.03 | 0.31 | 40.0 | 40.0 |
| **46** | Linguistic | Language Modeling | Percentile 10 of 4gram probabilities | 0.21 | 0.2 | 0.33 | 0.74 | 40.0 | 40.0 |
| **47** | Linguistic | Language Modeling | Percentile 50 of 1gram probabilities | 0.0 | 0.01 | -1.4 | 0.17 | 40.0 | 40.0 |
| **48** | Linguistic | Language Modeling | Percentile 50 of 2gram probabilities | 0.03 | 0.03 | -0.15 | 0.88 | 40.0 | 40.0 |
| **49** | Linguistic | Language Modeling | Percentile 50 of 3gram probabilities | 0.21 | 0.15 | 2.34 | 0.02 | 40.0 | 40.0 |
| **50** | Linguistic | Language Modeling | Percentile 50 of 4gram probabilities | 0.46 | 0.5 | -0.76 | 0.45 | 40.0 | 40.0 |
| **51** | Linguistic | Language Modeling | Percentile 90 of 1gram probabilities | 0.02 | 0.01 | 0.77 | 0.44 | 40.0 | 40.0 |
| **52** | Linguistic | Language Modeling | Percentile 90 of 2gram probabilities | 0.15 | 0.14 | 0.52 | 0.61 | 40.0 | 40.0 |
| **53** | Linguistic | Language Modeling | Percentile 90 of 3gram probabilities | 0.45 | 0.41 | 1.13 | 0.26 | 40.0 | 40.0 |
| **54** | Linguistic | Language Modeling | Percentile 90 of 4gram probabilities | 0.69 | 0.76 | -1.26 | 0.21 | 40.0 | 40.0 |
| **55** | Linguistic | Language Modeling | Perplexity of the entire sample | 68.24 | 60.61 | 0.44 | 0.66 | 40.0 | 40.0 |
| **56** | Linguistic | Repetitiveness | Lemmas repeated one word apart | 0.01 | 0.01 | 0.96 | 0.34 | 40.0 | 40.0 |
| **57** | Linguistic | Repetitiveness | Lemmas repeated two words apart | 0.03 | 0.03 | -0.16 | 0.87 | 40.0 | 40.0 |
| **58** | Linguistic | Repetitiveness | Lemmas repeated zero words apart | 0.0 | 0.0 | -1.2 | 0.24 | 40.0 | 40.0 |
| **59** | Linguistic | Repetitiveness | Part-of-speech tags repeated one word apart | 0.09 | 0.05 | 2.37 | 0.02 | 40.0 | 40.0 |
| **60** | Linguistic | Repetitiveness | Part-of-speech tags repeated two word apart | 0.13 | 0.13 | 0.22 | 0.82 | 40.0 | 40.0 |
| **61** | Linguistic | Repetitiveness | Part-of-speech tags repeated zero words apart | 0.05 | 0.03 | 1.0 | 0.32 | 40.0 | 40.0 |
| **62** | Linguistic | Semantic | Max distance to ICU "apron" | 0.39 | 0.4 | -0.34 | 0.73 | 40.0 | 40.0 |
| **63** | Linguistic | Semantic | Max distance to ICU "asking" | 0.59 | 0.62 | -1.57 | 0.12 | 40.0 | 40.0 |
| **64** | Linguistic | Semantic | Max distance to ICU "boy" | 0.89 | 0.9 | -0.24 | 0.81 | 40.0 | 40.0 |
| **65** | Linguistic | Semantic | Max distance to ICU "cabinet" | 0.52 | 0.54 | -0.73 | 0.47 | 40.0 | 40.0 |
| **66** | Linguistic | Semantic | Max distance to ICU "cookie" | 0.88 | 1.0 | -3.09 | 0.0 | 40.0 | 40.0 |
| **67** | Linguistic | Semantic | Max distance to ICU "counter" | 0.5 | 0.49 | 0.32 | 0.75 | 40.0 | 40.0 |
| **68** | Linguistic | Semantic | Max distance to ICU "cupboard" | 0.54 | 0.53 | 0.76 | 0.45 | 40.0 | 40.0 |
| **69** | Linguistic | Semantic | Max distance to ICU "curtains" | 0.48 | 0.44 | 1.5 | 0.14 | 40.0 | 40.0 |
| **70** | Linguistic | Semantic | Max distance to ICU "dishcloth" | 0.3 | 0.31 | -0.96 | 0.34 | 40.0 | 40.0 |
| **71** | Linguistic | Semantic | Max distance to ICU "dishes" | 0.74 | 0.77 | -0.84 | 0.41 | 40.0 | 40.0 |
| **72** | Linguistic | Semantic | Max distance to ICU "drying" | 0.61 | 0.6 | 0.2 | 0.84 | 40.0 | 40.0 |
| **73** | Linguistic | Semantic | Max distance to ICU "exterior" | 0.42 | 0.4 | 1.61 | 0.11 | 40.0 | 40.0 |
| **74** | Linguistic | Semantic | Max distance to ICU "falling" | 0.61 | 0.59 | 0.65 | 0.51 | 40.0 | 40.0 |
| **75** | Linguistic | Semantic | Max distance to ICU "faucet" | 0.61 | 0.64 | -0.88 | 0.38 | 40.0 | 40.0 |
| **76** | Linguistic | Semantic | Max distance to ICU "floor" | 0.79 | 0.72 | 1.21 | 0.23 | 40.0 | 40.0 |
| **77** | Linguistic | Semantic | Max distance to ICU "girl" | 0.87 | 0.87 | 0.12 | 0.9 | 40.0 | 40.0 |
| **78** | Linguistic | Semantic | Max distance to ICU "jar" | 0.78 | 0.79 | -0.03 | 0.98 | 40.0 | 40.0 |
| **79** | Linguistic | Semantic | Max distance to ICU "kitchen" | 0.68 | 0.64 | 1.07 | 0.29 | 40.0 | 40.0 |
| **80** | Linguistic | Semantic | Max distance to ICU "mother" | 0.87 | 0.89 | -0.66 | 0.51 | 40.0 | 40.0 |
| **81** | Linguistic | Semantic | Max distance to ICU "overflowing" | 0.45 | 0.48 | -0.94 | 0.35 | 40.0 | 40.0 |
| **82** | Linguistic | Semantic | Max distance to ICU "plate" | 0.59 | 0.59 | -0.01 | 0.99 | 40.0 | 40.0 |
| **83** | Linguistic | Semantic | Max distance to ICU "sink" | 0.85 | 0.94 | -1.9 | 0.06 | 40.0 | 40.0 |
| **84** | Linguistic | Semantic | Max distance to ICU "stealing" | 0.43 | 0.45 | -0.43 | 0.67 | 40.0 | 40.0 |
| **85** | Linguistic | Semantic | Max distance to ICU "stool" | 0.8 | 0.88 | -1.45 | 0.15 | 40.0 | 40.0 |
| **86** | Linguistic | Semantic | Max distance to ICU "washing" | 0.67 | 0.64 | 0.74 | 0.46 | 40.0 | 40.0 |
| **87** | Linguistic | Semantic | Max distance to ICU "water" | 0.89 | 0.76 | 2.59 | 0.01 | 40.0 | 40.0 |
| **88** | Linguistic | Semantic | Max distance to ICU "window" | 0.59 | 0.53 | 1.15 | 0.25 | 40.0 | 40.0 |
| **89** | Linguistic | Semantic | Max distance to ICU "woman" | 0.76 | 0.78 | -0.97 | 0.34 | 40.0 | 40.0 |
| **90** | Linguistic | Syntactic | Number of Dependency label sequence (advmod, root) normalized by total number of sentences | 0.08 | 0.03 | 1.37 | 0.18 | 40.0 | 40.0 |
| **91** | Linguistic | Syntactic | Number of dependency label advmod normalized by total nubmer of words | 0.01 | 0.01 | 0.93 | 0.35 | 40.0 | 40.0 |
| **92** | Linguistic | Syntactic | Number of dependency label appos normalized by total nubmer of words | 0.0 | 0.0 | -0.23 | 0.82 | 40.0 | 40.0 |
| **93** | Linguistic | Syntactic | Number of dependency label aux normalized by total nubmer of words | 0.06 | 0.08 | -1.91 | 0.06 | 40.0 | 40.0 |
| **94** | Linguistic | Syntactic | Number of dependency label case normalized by total nubmer of words | 0.11 | 0.09 | 1.98 | 0.05 | 40.0 | 40.0 |
| **95** | Linguistic | Syntactic | Number of dependency label dep normalized by total nubmer of words | 0.01 | 0.01 | -0.16 | 0.87 | 40.0 | 40.0 |
| **96** | Linguistic | Syntactic | Number of dependency label det normalized by total nubmer of words | 0.11 | 0.13 | -1.15 | 0.25 | 40.0 | 40.0 |
| **97** | Linguistic | Syntactic | Number of dependency label nmod normalized by total nubmer of words | 0.1 | 0.08 | 1.85 | 0.07 | 40.0 | 40.0 |
| **98** | Linguistic | Syntactic | Number of dependency label sequence (aux, root) normalized by total number of sentences | 0.41 | 0.58 | -2.25 | 0.03 | 40.0 | 40.0 |
| **99** | Linguistic | Syntactic | Number of dependency label sequence (case, nmod) normalized by total number of sentences | 1.09 | 0.97 | 0.64 | 0.53 | 40.0 | 40.0 |
| **100** | Linguistic | Syntactic | Number of dependency label sequence (det, nmod) normalized by total number of sentences | 0.49 | 0.5 | -0.08 | 0.94 | 40.0 | 40.0 |
| **101** | Linguistic | Syntactic | Number of dependency label sequence (det, nsubj) normalized by total number of sentences | 0.4 | 0.65 | -1.81 | 0.07 | 40.0 | 40.0 |
| **102** | Linguistic | Syntactic | Number of dependency label sequence (nmod:poss, nmod) normalized by total number of sentences | 0.03 | 0.04 | -0.45 | 0.65 | 40.0 | 40.0 |
| **103** | Linguistic | Syntactic | Number of dependency label sequence (nsubj, root) normalized by total number of sentences | 0.65 | 0.82 | -2.15 | 0.03 | 40.0 | 40.0 |
| **104** | Linguistic | Syntactic | Number of determiners normalized by number of words | 0.02 | 0.04 | -2.08 | 0.04 | 40.0 | 40.0 |
| **105** | Linguistic | Syntactic | Number of gerunds normalized by number of words | 0.0 | 0.01 | -2.53 | 0.01 | 40.0 | 40.0 |
| **106** | Linguistic | Syntactic | Number of nouns normalized by number of words | 0.35 | 0.33 | 1.19 | 0.24 | 40.0 | 40.0 |
| **107** | Linguistic | Syntactic | Number of prepositions normalized by number of words | 0.12 | 0.09 | 2.56 | 0.01 | 40.0 | 40.0 |
| **108** | Linguistic | Syntactic | Number of verbs normalized by number of words | 0.24 | 0.28 | -3.11 | 0.0 | 40.0 | 40.0 |
| **109** | Linguistic | Syntactic | Part-of-speech tag counts of JJ normalized by sentences | 0.33 | 0.21 | 1.44 | 0.15 | 40.0 | 40.0 |
| **110** | Linguistic | Syntactic | Part-of-speech tag counts of NN normalized by sentences | 2.54 | 2.36 | 0.48 | 0.64 | 40.0 | 40.0 |
| **111** | Linguistic | Syntactic | Part-of-speech tag counts of VBD normalized by sentences | 0.12 | 0.12 | -0.08 | 0.93 | 40.0 | 40.0 |
| **112** | Linguistic | Syntactic | Part-of-speech tag counts of VBG normalized by sentences | 1.27 | 1.35 | -0.49 | 0.63 | 40.0 | 40.0 |
| **113** | Linguistic | Syntactic | Part-of-speech tag counts of VBZ normalized by sentences | 0.93 | 1.1 | -0.97 | 0.33 | 40.0 | 40.0 |
| **114** | Linguistic | Verbosity - Diversity | Median number of characters in words | 3.45 | 3.3 | 1.42 | 0.16 | 40.0 | 40.0 |
| **115** | Linguistic | Verbosity - Diversity | Total number of charachers | 157.4 | 151.68 | 0.38 | 0.7 | 40.0 | 40.0 |
| **116** | Linguistic | Verbosity - Diversity | Total number of sentences | 3.85 | 3.75 | 0.23 | 0.82 | 40.0 | 40.0 |
| **117** | Linguistic | Verbosity - Diversity | Total number of unique words | 27.77 | 27.27 | 0.23 | 0.82 | 40.0 | 40.0 |
| **118** | Linguistic | Verbosity - Diversity | Type-token ratio | 0.78 | 0.77 | 0.41 | 0.68 | 40.0 | 40.0 |
| **119** | Linguistic | Verbosity - Diversity - Repetitiveness | Total number words | 37.62 | 36.65 | 0.24 | 0.81 | 40.0 | 40.0 |
| **120** | Linguistic | Word Frequency Effect | Median word frequencies | 814.55 | 1196.11 | -1.45 | 0.15 | 40.0 | 40.0 |
| **121** | Linguistic | Writing | Misspelling exists | 7870.02 | 4790.45 | 2.05 | 0.04 | 40.0 | 40.0 |
| **122** | Linguistic | Writing | Number of commas | 0.0 | 0.01 | -2.16 | 0.03 | 40.0 | 40.0 |
| **123** | Linguistic | Writing | Number of misspelled words normalized by total number of words | 0.03 | 0.02 | 1.12 | 0.27 | 40.0 | 40.0 |
| **124** | Linguistic | Writing | Uppercased letter at the beginning of a line, normalized by total number of lines | 0.61 | 0.71 | -1.26 | 0.21 | 40.0 | 40.0 |

1. Snowdon DA. Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life. Findings from the Nun Study. JAMA J Am Med Assoc. 1996;275:528–532.

2. Engelman M, Agree EM, Meoni LA, Klag MJ. Propositional Density and Cognitive Function in Later Life: Findings From the Precursors Study. J Gerontol Ser B. 2010;65B:706–711.

3. Pakhomov SVS, Hemmy LS. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. Cortex. 2014;55:97–106.

4. Orimaye SO, Wong JS-M, Golden KJ. Learning Predictive Linguistic Features for Alzheimer’s Disease and related Dementias using Verbal Utterances. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics; 2014.

5. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer’s Disease in Narrative Speech. J Alzheimer’s Dis. 2015;49:407–422.

6. Orimaye SO, Wong JS-M, Golden KJ, Wong CP, Soyiri IN. Predicting probable Alzheimer’s disease using linguistic deficits and biomarkers. BMC Bioinformatics. 2017;18.

7. Guerrero JM, Rincón M, Peraita H, Martínez-Tomás R. Diagnosis of Cognitive Impairment Compatible with Early Diagnosis of Alzheimer’s Disease. Methods Inf Med. 2016;55:42–49.

8. Aluísio S, Cunha A, Scarton C. Evaluating Progression of Alzheimer’s Disease by Regression and Classification Methods in a Narrative Language Test in Portuguese. In: Lecture Notes in Computer Science. Springer International Publishing; 2016. p. 109–114.

9. Kato S, Endo H, Homma A, Sakuma T, Watanabe K. Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2013.

10. Tröger J, Linz N, Alexandersson J, König A, Robert P. Automated speech-based screening for alzheimer\textquotesingles disease in a care service scenario. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare - PervasiveHealth \textquotesingle17. ACM Press; 2017.

11. Tröger J, Linz N, König A, Robert P, Alexandersson J. Telephone-based Dementia Screening I. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare - PervasiveHealth 18. ACM Press; 2018.

12. Kato S, Homma A, Sakuma T, Nakamura M. Detection of mild Alzheimer\textquotesingles disease and mild cognitive impairment from elderly speech: Binary discrimination using logistic regression. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015.

13. Gosztolya G, Tóth L, Grósz T, Vincze V, Hoffmann I, Szatlóczki G, et al. Detecting Mild Cognitive Impairment from Spontaneous Speech by Correlation-Based Phonetic Feature Selection. In: Interspeech 2016. ISCA; 2016.

14. Alhanai T, Au R, Glass J. Spoken language biomarkers for detecting cognitive impairment. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2017.

15. Al-Hameed S, Benaissa M, Christensen H. Detecting and Predicting Alzheimer\textquotesingles Disease Severity in Longitudinal Acoustic Data. In: Proceedings of the International Conference on Bioinformatics Research and Applications 2017 - ICBRA 2017. ACM Press; 2017.

16. Konig A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, et al. Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People. Curr Alzheimer Res. 2018;15.

17. Eyigöz E, Cecchi GA, Tejwani R. Predicting Cognitive Impairments with a Mobile Application. In: ICAART (2). 2018. p. 683–692.

18. Yancheva M, Fraser K, Rudzicz F. Using linguistic features longitudinally to predict clinical scores for Alzheimer\textquotesingles disease and related dementias. In: Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. Association for Computational Linguistics; 2015.

19. Clark DG, Kapur P, Geldmacher DS, Brockington JC, Harrell L, DeRamus TP, et al. Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. Cortex. 2014;55:202–218.

20. Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, et al. Identification of Conversion from Mild Cognitive Impairment to Alzheimer\textquotesingles Disease Using Multivariate Predictors. PLoS ONE. 2011;6:e21896.

21. Pereira T, Lemos L, Cardoso S, Silva D, Rodrigues A, Santana I, et al. Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. BMC Med Inform Decis Mak. 2017;17.

22. Preische O, Schultz S, Apel A, Kuhle J, Kaeser S, Barro C, et al. Dominantly Inherited Alzheimer Network. Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer’s disease. Nat Med. 2019;25:277–283.

23. Lin M, Gong P, Yang T, Ye J, Albin RL, Dodge HH. Big Data Analytical Approaches to the NACC Dataset. Alzheimer Dis Assoc Disord. 2018;32:18–27.

24. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data. NPJ Digit Med. 2020;3:1–7.

25. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology. 2019;290:456–464.

26. Pakhomov SVS, Eberly L, Knopman D. Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. Neuropsychologia. 2016;89:42–56.

27. Pakhomov SV, Eberly LE, Knopman DS. Recurrent perseverations on semantic verbal fluency tasks as an early marker of cognitive impairment. J Clin Exp Neuropsychol. 2018;40:832–840.

28. Raoux N, Amieva H, Le Goff M, Auriacombe S, Carcaillon L, Letenneur L, et al. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer’s disease subjects: Results from the PAQUID longitudinal study. Cortex. 2008;44:1188–1196.

29. Silverman JM, Li G, Zaccario ML, Smith CJ, Schmeidler J, Mohs RC, et al. Patterns of risk in first-degree relatives of patients with Alzheimer’s disease. Arch Gen Psychiatry. 1994;51:577–586.

30. Silverman JM, Smith CJ, Marin DB, Mohs RC, Propper CB. Familial patterns of risk in very late-onset Alzheimer disease. Arch Gen Psychiatry. 2003;60:190–197.

31. Silverman JM, Ciresi G, Smith CJ, Marin DB, Schnaider-Beeri M. Variability of familial risk of Alzheimer disease across the late life span. Arch Gen Psychiatry. 2005;62:565–573.

32. West DB, others. Introduction to graph theory. Vol. 2. Prentice hall Upper Saddle River; 2001.

33. Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology. 2000;14:71–91.

34. Kavé G, Dassa A. Severity of Alzheimer’s disease and language features in picture descriptions. Aphasiology. 2017;32:27–40.

35. Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer’s disease: A review of picture description tasks. J Clin Exp Neuropsychol. 2018;1–23.

36. Cullen B, Coen RF, Lynch CA, Cunningham CJ, Coakley D, Robertson IH, et al. Repetitive behaviour in Alzheimer\textquotesingles disease: description, correlates and functions. Int J Geriatr Psychiatry. 2005;20:686–693.

37. Barton S, Findlay D, Blake RA. The management of inappropriate vocalisation in dementia: a hierarchical approach. Int J Geriatr Psychiatry. 2005;20:1180–1186.

38. Lira JO de, Ortiz KZ, Campanha AC, Bertolucci PHF, Minett TSC. Microlinguistic aspects of the oral narrative in patients with Alzheimer\textquotesingles disease. Int Psychogeriatr. 2010;23:404–412.

39. Rapcsak SZ, Arthur SA, Bliklen DA, Rubens AB. Lexical Agraphia in Alzheimer\textquotesingles Disease. Arch Neurol. 1989;46:65–68.

40. Fukui T, Lee E. Progressive agraphia can be a harbinger of degenerative dementia. Brain Lang. 2008;104:201–210.

41. Giles E, Patterson K, Hodges JR. Performance on the Boston Cookie theft picture description task in patients with early dementia of the Alzheimer\textquotesingles type: Missing information. Aphasiology. 1996;10:395–408.

42. CUETOS F, ARANGO-LASPRILLA JC, URIBE C, VALENCIA C, LOPERA F. Linguistic changes in verbal expression: A preclinical marker of Alzheimer\textquotesingles disease. J Int Neuropsychol Soc. 2007;13.

43. Ahmed S, Haigh A-MF, Jager CA de, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer’s disease. Brain. 2013;136:3727–3737.

44. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–1543.

45. Yancheva M, Rudzicz F. Vector-space topic models for detecting Alzheimer\textquotesingles disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2016.

46. Ahmed S, Jager CA de, Haigh A-M, Garrard P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer\textquotesingles disease. Neuropsychology. 2013;27:79–85.

47. Yancheva M, Rudzicz F. Vector-space topic models for detecting Alzheimer\textquotesingles disease. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2016.

48. Turner A, Greene E. The construction and use of a propositional text base. Institute for the Study of Intellectual Behavior, University of Colorado …; 1977.

49. Chand V, Baynes K, Bonnici L, Farias ST. Analysis of idea density (AID): A manual. Univ Calif Davis. 2010;

50. Sirts K, Piguet O, Johnson M. Idea density for predicting Alzheimer\textquotesingles disease from transcribed speech. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics; 2017.

51. Brown C, Snodgrass T, Kemper SJ, Herman R, Covington MA. Automatic measurement of propositional idea density from part-of-speech tagging. Behav Res Methods. 2008;40:540–545.

52. Rosenzweig MR, Postman L. Frequency of usage and the perception of words. Science. 1958;127:263–266.

53. Howes D. On the relation between the intelligibility and frequency of occurrence of English words. J Acoust Soc Am. 1957;29:296–305.

54. Montanes P, Goldblum MC, Boller F. The naming impairment of living and nonliving items in Alzheimer\textquotesingles disease. J Int Neuropsychol Soc. 1995;1:39.

55. Bird H, Ralph MAL, Patterson K, Hodges JR. The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. Brain Lang. 2000;73:17–49.

56. Blazely AM, Coltheart M, Casey BJ. Semantic impairment with and without surface dyslexia: Implications for models of reading. Cogn Neuropsychol. 2005;22:695–717.

57. Bridges KA, Sidtis DVL. Formulaic language in Alzheimer\textquotesingles disease. Aphasiology. 2013;27:799–810.

58. Stolcke A. SRILM-an extensible language modeling toolkit. In: Seventh international conference on spoken language processing. 2002.

59. Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer’s disease: description of study cohort and accuracy of diagnosis. Arch Neurol. 1994;51:585–594.

60. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Comput Speech Lang. 1999;13:359–394.

61. Kempler D, Almor A, Tyler LK, Andersen ES, MacDonald MC. Sentence Comprehension Deficits in Alzheimer\textquotesingles Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. Brain Lang. 1998;64:297–316.

62. Lyons K, Kemper S, Labarge E, Ferraro FR, Balota D, Storandt M. Oral language and alzheimer\textquotesingles disease: A reduction in syntactic complexity. Aging Neuropsychol Cogn. 1994;1:271–281.

63. Bickel C, Pantel J, Eysenbach K, SchrÃ¶der J. Syntactic Comprehension Deficits in Alzheimer\textquotesingles Disease. Brain Lang. 2000;71:432–448.

64. Kemper S. On the Preservation of Syntax in Alzheimer\textquotesingles Disease. Arch Neurol. 1993;50:81.

65. Croisile B, Brabant M-J, Carmoi T, Lepage Y, Aimard G, Trillet M. Comparison between Oral and Written Spelling in Alzheimer\textquotesingles Disease. Brain Lang. 1996;54:361–387.

66. Waters GS, Rochon E, Caplan D. Task Demands and Sentence Comprehension in Patients with Dementia of the Alzheimer\textquotesingles Type. Brain Lang. 1998;62:361–397.

67. Roark B, Mitchell M, Hollingshead K. Syntactic complexity measures for detecting mild cognitive impairment. In: Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP \textquotesingle07. Association for Computational Linguistics; 2007.

68. Eyigöz E, Pietrowicz M, Agurto C, Orozco J-R, Garcia A, Skodda S, et al. Dependency Analysis of Spoken Language for Assessment of Neurological Disorders. In: LREC workshop: RaPID-3. 2020.

69. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer’s Disease in Narrative Speech. J Alzheimer’s Dis. 2015;49:407–422.

70. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.

71. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc J Atmospheric Sci Appl Meteorol Phys Oceanogr. 2002;128:2145–2166.