Published in final edited form as:

Appl Neuropsychol Adult. 2022 ; 29(5): 1250–1257. doi:10.1080/23279095.2020.1864733.

Automated Assessment of Speech Production and Prediction of MCI in Older Adults

Victoria Sanborn, M.A.[a,*], **Rachel Ostrand, Ph.D.**[b], **Jeffrey Ciesla, Ph.D.**[a], **John Gunstad, Ph.D.**[a,c]

[a]Department of Psychological Sciences, Kent State University, Kent, OH, U.S.

[b]Department of Healthcare & Life Sciences, IBM Research, Yorktown Heights, NY, U.S.

[c]Brain Health Research Institute, Kent State University, Kent, OH U.S.

Abstract

The population of older adults is growing dramatically and, with it comes increased prevalence of neurological disorders, including Alzheimer's disease (AD). Though existing cognitive screening tests can aid early detection of cognitive decline, these methods are limited in their sensitivity and require trained administrators. The current study sought to determine whether it is possible to identify persons with mild cognitive impairment (MCI) using automated analysis of spontaneous speech. Participants completed a brief neuropsychological test battery and a spontaneous speech task. MCI was classified using established research criteria, and lexical-semantic features were calculated from spontaneous speech. Logistic regression analyses compared the predictive ability of a commonly-used cognitive screening instrument (the Modified Mini Mental Status Exam, 3MS) and speech indices for MCI classification. Testing against constant-only logistic regression models showed that both the 3MS [$\chi^2(1)=6.18$, p=0.013; AIC=41.46] and speech indices [$\chi^2(16)=32.42$, p= 0.009; AIC=108.41] were able to predict MCI status. Follow-up testing revealed the full speech model better predicted MCI status than did 3MS (p=.049). In combination, the current findings suggest that spontaneous speech may have value as a potential screening measure for identification of cognitive deficits, though confirmation is needed in larger, prospective studies.

Keywords

Alzheimer's disease; cognitive dysfunction; speech; cognition; aged

Introduction

The population of older adults is growing dramatically worldwide. By 2030, the population of adults over age 65 is projected to reach 70 million in the United States (Ortman, Velkoff, & Hogan, 2014) and 1 billion globally (He, Goodkind, & Kowal, 2016). This

[*]Corresponding author: Victoria Sanborn, M.A.; 600 Hilltop Drive, 144 Kent Hall, Kent State University Kent, OH 44240; vsanborn@kent.edu.

This work was conducted jointly at Kent State University and IBM Research.

societal trend is likely to increase the prevalence of neurological conditions. For example, the number of persons with Alzheimer's disease (AD) is expected to triple by 2050 and produce an estimated $1.1 trillion in annual health care costs in the U.S. alone (Hebert et al., 2013; Alzheimer's Association, 2018). As such, early detection of AD and other forms of pathological cognitive aging is urgently needed.

Numerous brief cognitive screening instruments have been developed, though these paper-and-pencil measures are not routinely administered in many settings due to a variety of practical barriers (Boustani et al., 2003; Khachaturian et al., 2009). Further, screeners that are commonly administered often show limited ability to identify those persons with more subtle cognitive deficits (Bradford et al., 2009; Chodosh et al., 2004; Valcour et al., 2000; Behrman, Valkanova, & Allan, 2017; Mitchell, 2009). Given these limitations, alternative approaches for monitoring cognitive function are needed.

It may be possible to utilize speech analysis to assist in the detection of early cognitive decline. Speech production is a complex neural activity which draws upon many interacting neural systems, including memory and executive function. Clinical and case studies have shown that changes in speech production are common in persons with mild cognitive impairment (MCI) and AD (e.g., Bayles, Tomoeda, & Trosset, 1992; Henry, Crawford, & Phillips, 2004; Nicholas, Obler, Albert, & Helm-Estabrooks, 1985). For example, persons with AD often exhibit empty speech (i.e., producing vague rather than specific words such as "thing" in place of "armchair") and use higher frequency (i.e., more common) words, which are easier to access from semantic memory (Nicholas, Obler, Albert, & Helm-Estabrooks, 1985; Kavé & Dassa, 2018). Consistent with this approach, traditional neuropsychological tests of language function (e.g., Animal Naming, Boston Naming Test) have long been used to help diagnose AD and identify those persons at risk for future conversion to AD (e.g., Henry, Crawford, & Phillips, 2004; Eastman et al., 2013; Pravatà, Tavernier, Parker, Vavro, Mintzer, & Spampinato, 2016; Pakhomov, Eberly, & Knopman, 2018; Blackwell, Sahakian, Vesey, Semple, Robbins, & Hodges, 2004).

Recent research suggests that spontaneous speech may be even more sensitive to AD risk than are traditional neuropsychological tests of language function (Bayles et al., 1992; Nicholas et al., 1985; Bucks, Singh, Cuerden, & Wilcock, 2000; Fraser, Meltzer, & Rudzicz, 2016; Giles, Patterson, & Hodges, 1996; Kavé & Dassa, 2018; Toth et al., 2018; Meilan et al., 2014; Meilan et al., 2018; Lopez-de-Ipina, 2018; Konig, et al., 2017; Misiewicz, et al., 2017). Spontaneous speech can be collected by asking individuals to describe a picture (such as the Cookie Theft; [Goodglass H, Kaplan E. The Assessment of Aphasia and Related Disorders. 2nd ed. Lea & Febiger; 1983.]), engage in a semi-structured guided interview with the examiner, or retell a well-known story. Responses are audio-recorded and transcribed by human listeners. Linguistic characteristics can then be computed using the transcriptions and recordings, and divided into different linguistic levels, including low-level acoustic and temporal (e.g., speech rate, duration of pauses and hesitations), lexical-semantic (e.g., word choice, word finding difficulties, repetitions, empty speech), morphosyntactic (e.g., syntactic structures and inflection errors), and discourse/pragmatic elements (e.g., cohesion, diversity of word choice) (Boschi et al., 2017; Slegers et al., 2018). Recent reviews have characterized the many changes in spontaneous speech observed in persons with AD, including frequent hesitations, semantic and lexical errors, repetitions, greater inflectional errors, and reduced referential and temporal cohesion (Boschi et al., 2017; Filiou et al, 2019; Slegers et al., 2018).

Technological advances have greatly enhanced the potential of utilizing speech analysis to help monitor cognitive function over time. Historically, analysis of spontaneous speech required substantial human labor and linguistic training to transcribe, code, and analyze speech. New approaches such as automatic speech recognition and machine learning techniques, as well as automated ways to measure various linguistic features from an audio recording or transcript have automated many of these processes while maintaining high levels of accuracy. One recent study automatically calculated a suite of lexical-semantic linguistic features from several spontaneous speech tasks and found that these linguistic features were predictive of both current and future cognitive test performance in older adults without dementia (Ostrand & Gunstad, 2020).

When combined with the many practical advantages of speech-based screening relative to existing methods (e.g., repeatability, scalability, and self-administration), such findings encourage examination of the utility of spontaneous speech as a method for monitoring cognitive status over time. The current study sought to determine whether a spontaneous speech task could be used to predict MCI in a sample of community-dwelling older adults. We chose to focus on lexical-semantic aspects of speech, as decline in these abilities is frequently observed in persons with pathological cognitive decline (e.g., word finding difficulties, vague or empty speech) and techniques for automated generation of these speech indices has been previously validated.

Materials and Methods

Participants

A total of 90 individuals completed the study protocol, though two were excluded prior to data analyses due to missing data that precluded determination of MCI status (years of education and Digit Span). Data from the remaining 88 participants were analyzed to test study hypotheses. See Tables 1 and 2 for sample characteristics.

Speech tasks and indices

To elicit a spontaneous speech monologue sample from participants, the experimenter provided a picture book of the fairy tale *Cinderella* with the words removed. The participant looked through the pictures to remind themselves of the story, gave the book back to the experimenter, and then retold the story from memory (following Saffran, Berndt, & Schwartz, 1989; see also MacWhinney, Fromm, Holland, Forbes, & Wright 2010). This task has been used in past work to assess people with aphasia as well as healthy, non-aphasic controls, and has been shown to be sensitive to lexical patterns and morphosyntactic control in persons with intact cognitive function (MacWhinney et al., 2010; Fromm et al., 2017). Several factors guided the selection of this speech elicitation task. First, it provides a middle ground of task constraint, in between other common speech elicitation tasks – higher constraint than an open-ended interview question, as it gives participants some amount of semantic context from the storyline, but lower constraint than picture description, where the participant is bound by the content of the picture. As the goal of current study is to investigate whether individual participants' variability in linguistic measures can account for their variability in cognitive status, a speech task which allows for greater variability in behavior between participants may be a more effective predictor of cognitive status. Additionally, a story-retelling task imposes higher memory demands than does a picture description task, as the participant must remember the storyline without having external memory support from the visual cues to guide their speech and recall. As a result, retelling a story draws on not just semantic memory to retrieve appropriate words, but also episodic memory of the story itself as well as attentional and executive function controls to keep the thread of the story continuous and comprehensible. Similarly, retelling a story avoids overt labelling of nouns which may occur during a picture description task.

Responses were audio-recorded and later transcribed. Specifically, participants' speech samples were recorded using a Shure SM10A head-mounted, directional (cardioid) microphone, which isolates the participant's speech from the experimenter's voice and other background noise. Recordings were manually transcribed and time-stamped off-line by trained transcribers who were blind to the participant's cognitive status and were checked by a second trained transcriber.

A collection of lexical-semantic features of speech, based on those used in past work, were calculated automatically from the transcribed text using Python (version 2.7.17). Part-of-speech tags were computed using the Natural Language Toolkit (NLTK, version 3.2.1; Bird, Klein, & Loper, 2009) and the Penn Treebank tagset (Marcus, Santorini, & Marcinkiewicz, 1993). Lexical frequency indices were computed based on corpus data from the widely-used Switchboard and Fisher corpora (Godfrey & Holliman, 1993; Cieri, Graff, Kimball, Miller, & Walker, 2004; Cieri, Graff, Kimball, Miller, & Walker, 2005), which jointly comprise 24 million words. A description of each of the linguistic features is presented in Table 3.

Neuropsychological test battery

A brief battery was administered in a fixed order under the supervision of a licensed clinical neuropsychologist. Specific clinical tests included the 3MS (Teng & Chui, 1987), Hopkins Verbal Learning Test (Brandt & Benedict, 2001), Complex Figure Test (Meyers & Meyers, 1995; Berry, Allen, & Schmitt, 1991), Digit Span (Weschler, 2008), Trail Making Test A and B (Reitan, 1958), Frontal Assessment Battery (Dubois, Slachevsky, Litvan, & Pillon, 2000), Controlled Oral Word Association Test (Lezak, Howieson, & Loring, 2004), Animal Naming (Lezak et al., 2004), and Boston Naming Test – Short Form (Williams, Mack, & Henderson, 1989). The 3MS was chosen over other global cognitive screening tests, such as the Mini-Mental State Exam (MMSE), as it has been found to better identify persons with diagnosis of MCI and captures greater variability in performance reflecting cognitive domains most often associated with Alzheimer's disease (Van Patten, Britton, & Tremont, 2019). Normative values (i.e., t-scores) adjusting for age and education (Jones, Schinka, & Vanderploeg, 2002) were used to characterize test performance (see Table 1).

The 88 participants were classified into two groups based on established criteria (Jak et al., 2016): Intact (N = 62) vs. MCI (N = 26). Specifically, participants with two or more t-scores less than 40 in at least one cognitive domain on objective testing were identified as meeting criteria for MCI.

Procedure

After being evaluated for capacity and completing informed consent, participants completed a 75-minute session comprised of questionnaires, neuropsychological testing, and spontaneous speech tasks. At the completion of testing, participants were compensated for their time.

Data Analysis

All analyses were performed using IBM SPSS 25. After generating descriptive statistics to characterize the sample, chi-square and t-tests were used to compare persons with and without MCI. A series of logistic regressions were then used to identify the extent to which the 3MS and the automatically-calculated speech features from the Cinderella task could predict group status (MCI vs intact). In the regression using 3MS as a predictor, participants' 3MS score was t-scored, accounting for age and education, in order to promote consistency to its use in clinical settings. In the regression analysis using the linguistic features as predictors, all linguistic features were simultaneously entered into the logistic model as predictors in a single step. McNemar's test was used to compare the predictive ability of these two models. Finally, exploratory logistic regressions were conducted to further investigate the nature of the association between individual speech indices and MCI status.

Results

Sample characteristics

For the full sample, participants averaged 68.03 ± 7.90 years of age, 67.0% were female, and completed an average of 15.41 ± 2.56 years of education (see Table 1). There were no significant group differences in any demographic or medical characteristics between the participants identified as having intact cognitive function (n = 62) and those with MCI (n = 26) (all p > 0.05). As expected through operationalization of MCI using Jak criteria, Intact and MCI groups differed on most neuropsychological tests and many speech indices (see Table 2).

Of note, the groups differed on the 3MS [t(87) = 2.62, p = 0.01], though average 3MS t-scores for both groups fell within the normal range (Intact = 58.19 ± 6.23 ; MCI = 53.19 ± 11.64) and exhibited similar proportions of participants with t-scores less than 40 (Intact = 3.2%; MCI = 7.7%; χ^2 (1) = 0.83, p = 0.36).

Using 3MS t-scores to predict MCI status

A logistic regression analysis was performed with 3MS t-scores as the predictor variable and MCI status as the binary outcome variable. The overall proportion of variance accounted for was modest (Nagelkerke $R^2 = 0.10$), with the model showing sensitivity of 7.7% (i.e., 2 out of 26 MCI participants correctly classified as MCI) and specificity of 95.2% (i.e., 59 out of 62 intact participants correctly classified as intact). The unstandardized Beta weight for the predictor was $\beta = -0.72$ (SE=0.03, $Wald$=4.85, p=0.03) and estimated odds ratio indicated a reduced likelihood of falling into the MCI group by a factor of *Exp*(*β*) 0.931 (95% CI = 0.873 to 0.992) for each one unit increase in 3MS t score. This model results in an overall classification accuracy of 69.3%, which is significantly better than an intercept-only model [χ^2 (1) = 6.18, p = 0.013; AIC 41.46]. However, it is important to note that group sizes are unbalanced in the present sample, with 70.5% of participants in the Intact group and 29.5% of participants in the MCI group. Thus, although the model using 3MS as a predictor produced an overall prediction accuracy of 69.3%, this is actually *lower* than the accuracy that would be obtained by a model which assigned all participants to the majority class (in this case, Intact). As a result, although the comparison of this model against an intercept-only model was statistically significant, the model's lackluster performance when compared to chance highlights the need for novel approaches to rapidly assess cognitive status.

Using spontaneous speech indices to predict MCI status

A logistic regression was performed using the full set of speech indices jointly as predictors and MCI status as the binary outcome variable. See Table 4 for the output of the logistic regression. This model showed 50% sensitivity (i.e., 13 out of 26 persons with MCI correctly identified) and 91.9% specificity (i.e., 57 out of 62 intact persons correctly identified; Nagelkerke R^2 = 0.44). The model using the speech features as predictors explained significantly more variance than a constant-only model [χ^2(16) = 32.42, p = 0.009] with 79.5% accuracy.

An important point, however, is that many of the individual linguistic features are highly correlated with each other – for example, type-token ratio, Honoré's statistic, and Brunet's index are different but related ways of measuring lexical diversity. As a result, many of the predictors in this multiple regression are highly collinear with each other, making interpretation of the individual β weights from the multiple regression difficult (predictors which are highly collinear may result in β weights which are largely loaded onto Predictor A, Predictor B, or unpredictably split between the two). Although we report β weights for the multiple regression, they may not be meaningful in interpreting the contribution of individual linguistic features towards predicting MCI group status and thus should be considered with caution.

Therefore, to investigate the relationship between each individual linguistic feature and MCI status, separate binary logistic regression analyses were performed using each speech feature individually as a predictor of MCI status. Table 5 shows the outcome of each individual regression. Numerous indices showed significant improvement in prediction over the intercept-only model, with Nouns (Wald = 4.52, p =.002), Determiners (Wald = 7.78, p = .005) and Honoré's statistic (Wald = 7.94, p = .005) remaining so after correcting for multiple comparisons (Benjamini & Hochberg, 1995).

Comparing 3MS to Full Speech model

McNemar's test showed that the full model of linguistic features better predicted MCI group status than did the model using 3MS t-score as predictor (p = .049).

Exploratory Speech Models to Predict MCI Status

Three exploratory analyses were then conducted to examine the extent to which various combinations of speech indices could predict MCI status. The first utilized the three speech indices above that showed a significant independent relationship to MCI status after correction (i.e., Nouns, Determiners, and Honoré’s statistic). This model explained more variance than did the intercept only model [$\chi^2(3) = 13.28$, p = 0.004; Nagelkerke $R^2 = 0.20$] and exhibited 19.2% sensitivity and 93.5% specificity. However, McNemar’s test indicated this model did not differ from those using the 3MS (p = 0.75) or full collection of speech indices (p = 0.12) as predictors.

A second analysis sought to limit multicollinearity across speech indices. A linear regression was first performed after centering all values, and speech indices with an elevated variance inflation factor (VIF ≥ 5) were removed (Kutner, Nachtsheim, & Neter, 2004), specifically: Total words (VIF = 37.46), Filler words (VIF = 43.51), Definite article (VIF = 16.32), Pronouns (VIF = 5.65), Nouns (VIF = 20.03), Verbs (VIF = 19.35), Determiners (VIF = 31.34), Content words (VIF = 64.17), Type-Token ratio (VIF = 88.55), Brunet’s index (VIF = 106.76), and Filler rate (VIF = 36.29). Entering the five remaining indices (i.e., Empty words, Indefinite articles, Lexical frequency, Honoré’s statistic, and Speech rate) produced a model that performed significantly better than the intercept-only model [$\chi^2(5) = 12.68$, p = 0.027; Nagelkerke $R^2 = 0.19$], with a sensitivity of 23.1% and specificity of 93.5%, though did not differ from either the 3MS (p = 0.55) or full speech model (p = 0.21).

A final model was developed using speech indices that could be readily observed by clinicians, specifically Total words, Filler words, Empty words, Pronouns, Nouns, and Speech rate. This model was superior to the intercept-only model [$\chi^2(6) = 19.77$, p = 0.003; Nagelkerke $R^2 = 0.29$] and correctly identified 34.6% of persons with MCI and 90.3% of intact controls. Significance was largely driven by greater use of filler words (Exp (β) = 2.94, p = .02) and fewer nouns (Exp (β) = .15, p = .005) in persons with MCI. McNemar’s test indicated this model did not differ from the models using 3MS (p = .45) or the full collection of speech indices (p = .27) as predictors.

Discussion

The current study examined whether automatically-generated indices from spontaneous speech could be used to identify older adults meeting research criteria for MCI. Analyses showed that a combination of lexical-semantic speech features was somewhat better than the 3MS in predicting current cognitive status. Several aspects of these results warrant brief discussion.

Past work has shown that spontaneous speech indices are associated with neuropsychological test performance and are impaired in persons with conditions like AD (Ostrand & Gunstad, 2020; Pistono, Jucla, Bézy, Lemesle, Le Men, & Pariente, 2019; Pistono et al., 2016; Boschi et al., 2017; Bayles et al., 1992). This pattern is not surprising, as speech production is a complex neural process, particularly when considering the multiple cognitive processes needed to correctly retell a story without external cues. Features such as the use of fewer nouns and determiners, more filler words, and lower lexical diversity distinguished intact persons from those with MCI and may serve as an initial step toward the development of self-administered, ambulatory monitoring of cognitive status. For example, the current study used an automated approach when generating values for lexical-semantic features of speech, which dramatically reduces the time and effort needed for this task relative to traditional approaches. Continued advances in automatic speech recognition (ASR) and growing evidence for a link between cognitive function and other aspects of spontaneous speech (e.g., acoustic features, syntax, coherence; Boschi et al., 2017; Slegers et al., 2018) encourage further work in this area to develop automated approaches for clinical features such as circumlocutory speech or literal or semantic paraphasias. If successful, these tools could provide a method for broad screening for cognitive dysfunction through smart devices at home or even be modified to provide real-time information in clinical settings to assist with diagnosis.

However, prospective studies are much needed to better understand many aspects of spontaneous speech in older adults, including factors that contribute to normal between-subjects variability and identifying those linguistic indices that best distinguish normal aging from pathological conditions like AD. Socioeconomic and demographic factors such as age, race, ethnicity, education, premorbid lexicon, literacy levels, region, and bilingualism are well known to affect speech production (e.g., Kavé et al., 2009; Daller et al., 2003). In addition to these trait-like features, factors such as the identity of the listener, recent linguistic input, and concurrent memory load – among many others – can influence speech production. Similarly, the specific aspects of speech production which are affected by AD may change at various stages of the disease, as impairments in lexical access, in particular to higher-frequency and more specific words, appear to occur early in AD, whereas changes in syntactic production seem to emerge later (Davis & Maclagan, 2009; Ahmed, Haigh, de Jager, & Garrard, 2013; Snowdon, Kemper, Mortimer, Greiner, Wekstein, & Markesbery, 1996). Further, and as noted above, the present study focuses primarily on lexical-semantic features, and future investigation of other domains of speech (e.g., phonetic and phonological, morphosyntactic, and discourse levels) may provide additional insight and improve ability to detect MCI. Lastly, as noted in the present findings, determining which linguistic features of speech (taken in combination or individually) are associated with cognitive function is difficult and complex; despite differing outcomes in classifying MCI status, McNemar’s test showed no statistically significant differences in the models’ predictive abilities. Future research is needed to help clarify which outcomes may be clinically useful versus statistically significant and the benefits of utilizing continuous values from neuropsychological testing (c.f. Ostrand & Gunstad, 2020), though such studies will require larger samples and more diverse samples to draw conclusions.

The current study is limited in several important ways. The sample size was modest and study participants were highly educated, native English speakers, and resided within a single, largely monolingual region of the USA. These and other demographic factors are likely to affect performance on spontaneous speech indices as well as on neuropsychological test performance (Rosselli & Ardila, 2003; Saykin et al., 1995; Ardila & Rosselli, 1996). Additionally, research criteria for MCI was used rather than diagnosis through a comprehensive clinical evaluation. Though this approach is frequently used in past work and shows good predictive validity, it cannot replace a formal evaluation.

Conducting a comprehensive clinical evaluation in conjunction with spontaneous speech could provide important insight into underlying mechanisms, including the contribution of key neuroimaging markers (e.g., amyloid deposition, global vs. hippocampal atrophy) and other known risks for cognitive decline in older adults (e.g., APOE4; Nevler, Ash, Irwin, Liberman, & Grossman, 2019). Finally, additional work is needed to determine the most appropriate method for utilizing indices of spontaneous speech in neuropsychological research. Though the current study found a combination of lexical-semantic features was associated with cognitive status, future studies should evaluate the potential benefits of combining multiple features to represent components of speech (Cohen, Renshaw, Mitchell, & Kim, 2016; Cohen, Mitchell, Docherty, & Horan, 2016) as they may permit broad assessment of speech features while addressing statistical concerns.

In conclusion, the current study found that indices derived from spontaneous speech performed as well as a commonly used cognitive screening test frequently used in clinical settings in identifying older adults meeting research criteria for MCI. Additional studies are needed to further investigate automated speech analysis as a method to monitor cognitive decline in community settings.

Acknowledgments

Declaration of Interests

There are no conflicts of interest to report. Manuscript preparation was supported in part by NIH R01 AG065432 (JG and RO).

References

Ahmed S, Haigh A-MF, de Jager CA, & Garrard P (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer’s disease. Brain, 136(12). 3727–3737. doi:10.1093/brain/awt269 [PubMed: 24142144]

Alzheimer’s Association (2018). 2018 Alzheimer’s Disease Facts and Figures. Retrieved from www.alz.org/facts on 06/18/2018.

Ardila A, & Rosselli M (1996). Spontaneous language production and aging: sex and educational effects. International Journal of Neuroscience, 87(1-2), 71–78. 10.3109/00207459608990754 [PubMed: 8913820]

Bayles KA, Tomoeda CK, & Trosset MW (1992). Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. Brain and Language, 42(4), 454–472. 10.1016/0093-934X(92)90079-T [PubMed: 1377076]

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 (Statistical Methodology), 57(1), 289–300.

Behrman S, Valkanova V, & Allan C (2017). Diagnosing and managing mild cognitive impairment. Practitioner, 261, 17–20. PMID: 29120563

Berry DTR, Allen RS, & Schmitt FA (1991). Rey-Osterrieth complex figure: Psychometric characteristics in a geriatric sample. Clinical Neuropsychology, 5(2), 143–153. doi:10.1080/13854049108403298

Bird S, Klein E, & Loper E (2009). Natural Language Processing with Python. 1st ed. Sebastopol, CA: O’Reilly Media.

Blackwell AD, Sahakian BJ, Vesey R, Semple JM, Robbins TW, & Hodges JR (2004). Detecting dementia: Novel neuropsychological markers of preclinical Alzheimer’s disease. Dementia and Geriatric Cognitive Disorders, 17, 42–48. DOI:10.1159/000074081 [PubMed: 14560064]

Boschi V, Catricala E, Consonni M, Chesi C, Moro A, & Cappa SF (2017). Connected speech in neurodegenerative language disorders: A review. Frontiers in Psychology, 8, 269. 10.3389/fpsyg.2017.00269 [PubMed: 28321196]

Boustani M, Peterson B, Hanson L et al. (2003). Screening for dementia in primary care: a summary of the evidence for the U.S. Preventive Services Task Force. Annals of Internal Medicine, 138, 927–937. 10.7326/0003-4819-138-11-200306030-00015 [PubMed: 12779304]

Bradford A, Kunik M, Schulz P, et al. (2009). Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. Alzheimer’s Disease and Associated Disorders, 23, 306–314. Doi: 10.1097/WAD.0b013e3181a6bebc

Brandt J & Benedict RHB (2001) Hopkins Verbal Learning Test–Revised: Professional Manual. Lutz, FL: Psychological Assessment Resources.

Bucks RS, Singh S, Cuerden JM, & Wilcock GK (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. 10.1080/026870300401603

Chodosh J, Petitti D, Elliot M, et al. (2004). Physician recognition of cognitive impairment: evaluating the need for improvement. Journal of the American Geriatrics Society, 52, 1051–1059. 10.1111/j.1532-5415.2004.52301.x [PubMed: 15209641]

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

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

Cohen AS, Mitchell KR, Docherty NM, & Horan WP (2016). Vocal expression in schizophrenia: Less than meets the ear. Journal of Abnormal Psychology, 125(2), 299. [PubMed: 26854511]

Cohen AS, Renshaw TL, Mitchell KR, & Kim Y (2016). A psychometric investigation of “macroscopic” speech measures for clinical and psychological science. Behavior Research Methods, 48(2), 475–486. [PubMed: 25862539]

Daller H, Van Hout R, & Treffers-Daller J (2003). Lexical richness in the spontaneous speech of bilinguals. Applied Linguistics, 24(2), 197–222. 10.1093/applin/24.2.197

Davis BH & Maclagan M (2009). Examining pauses in Alzheimer’s discourse. American Journal of Alzheimer’s and Other Dementias, 24(2), 141–154. doi:10.1177/1533317508328138

Dubois B, Slachevsky A, Litvan I, & Pillon B (2000). The FAB: A frontal assessment battery at bedside. Neurology, 55(11):1621–1626. doi:10.1212/WNL.55.11.1621 [PubMed: 11113214]

Eastman JA, Hwang KS, Lazaris A, Chow N, Ramirez L, Babakchanian S, … & Apostolova LG (2013). Cortical thickness and semantic fluency in Alzheimer's disease and mild cognitive impairment. American journal of Alzheimer's disease (Columbia, Mo.), 1(2), 81. doi: 10.7726/ajad.2013.1006

Filiou R-P, Bier N, Slegers A, Houzé B, Belchior P, & Brambati SM (2019). Connected speech assessment in the early detection of Alzheimer’s disease and mild cognitive impairment: A scoping review. Aphasiology, 1–33. 10.1080/02687038.2019.1608502 [PubMed: 33012945]

Fraser KC, Meltzer JA, & Rudzicz F (2016). Linguistic features identify Alzheimer’s disease in narrative speech. Journal of Alzheimer's Disease, 49(2), 407–422. DOI: 10.3233/JAD-150520

Giles E, Patterson K, & Hodges JR (1996). Performance on the Boston Cookie Theft Picture Description Task in Patients with Early Dementia of the Alzheimer's type: Missing Information. Aphasiology, 10(4), 395–408. 10.1080/02687039608248419

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

He W, Goodkind D, & Kowal PR (2016). An aging world: 2015. International Population Reports. Retrieved on June 21, 2020.

Hebert L, Weuve J, Scherr P, et al. (2013). Alzheimer disease in the United State (2010-2050) estimated using the 2010 census. Neurology, 80, 1778–1783. 10.1212/WNL.0b013e31828726f5 [PubMed: 23390181]

Henry JD, Crawford JR, & Phillips LH (2004). Verbal fluency performance in dementia of the Alzheimer’s type: A meta-analysis. Neuropsychologia, 42(9), 1212–1222. 10.1016/j.neuropsychologia.2004.02.001 [PubMed: 15178173]

Jak AJ, Preis SR, Beiser AS, Seshadri S, Wolf PA, Bondi MW, & Au R (2016). Neuropsychological criteria for mild cognitive impairment and dementia risk in the Framingham Heart Study. Journal of the International Neuropsychological Society, 22(9), 937–943. [PubMed: 27029348]

Jones TG, Schinka JA, Vanderploeg RD, Small BJ, Graves AB, & Mortimer JA (2002) 3MS normative data for the elderly. Archives of Clinical Neuropsychology, 17(2), 171–177. doi:10.1016/s0887-6177(00)00108-6 [PubMed: 14589746]

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

Kavé G, Samuel-Enoch K, & Adiv S (2009). The association between age and the frequency of nouns selected for production. Psychology and Aging, 24(1), 17–27. 10.1037/a0014579 [PubMed: 19290734]

Khachaturian S, Camí J, Andrieu S, et al. (2009). Creating a transatlantic research enterprise for preventing Alzheimer's disease. Alzheimer's & Dementia, 5, 361–366. 10.1016/j.jalz.2009.05.158

Konig A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, & Robert PH (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. Current Alzheimer Research, 15(2), 120–129. 10.2174/1567205014666170829111942 [PubMed: 28847279]

Kutner M, Nachtsheim C, Neter J (2004). Applied linear statistical models. 4th. McGraw-Hill; Irwin.

Lezak MD, Howieson DB, & Loring DW (2004) Neuropsychological Assessment. 4th edition. Oxford; New York: Oxford University Press.

Lopez-de-Ipina K, Martinez-de-Lizarduy U, Calvo PM, Mekyska J, Beitia B, Barroso N, … & Ecay-Torres M (2018). Advances on automatic speech analysis for early detection of Alzheimer disease: a non-linear multi-task approach. Current Alzheimer Research, 15(2), 139–148. 10.2174/1567205014666171120143800 [PubMed: 29165084]

MacWhinney B, Fromm D, Holland A, Forbes M, & Wright H (2010). Automated analysis of the Cinderella story. Aphasiology, 24, 856–868. doi:10.1080/02687030903452632 [PubMed: 25067870]

Fromm D, Forbes M, Holland A, Dalton SG, Richardson J, & MacWhinney B (2017). Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff. American Journal of Speech-Language Pathology, 26(3), 762–768. [PubMed: 28505222]

Marcus MP, Santorini B, & Marcinkiewicz MA (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Comput Linguist, 19(2), 313–330.

Meilan J, et al. (2014). Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? Dementia and Geriatric Cognitive Disorders, 37, 327–334. 10.1159/000356726 [PubMed: 24481220]

Meilan J, et al. (2018). Voice markers of lexical access in mild cognitive impairment and Alzheimer's Disease. Current Alzheimer Research, 15, 111–119. 10.2174/1567205014666170829112439 [PubMed: 28847280]

Meyers JE & Meyers KR (1995). Rey Complex Figure Test and Recognition Trial -- Professional Manual. Odessa, FL: Psychological Assessment Resources.

Misiewicz S, Brickman AM, & Tosto G (2018). Prosodic impairment in dementia: review of the literature. Current Alzheimer Research, 15(2), 157–163. 10.2174/1567205014666171030115624 [PubMed: 29086698]

Mitchell A (2009). A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. Journal of Psychiatric Research, 43, 411–431. 10.1016/j.jpsychires.2008.04.014 [PubMed: 18579155]

Nevler N, Ash S, Irwin D, Liberman M, & Grossman M (2019). Prosodic Impairment as a Marker of apoE4 Status in logopenic variant Primary Progressive Aphasia with AD Pathology (P2. 1-032). Neurology, 92(15).

Nicholas M, Obler LK, Albert ML, & Helm-Estabrooks N (1985). Empty speech in Alzheimer's disease and fluent aphasia. Journal of Speech, Language, and Hearing Research, 28(3), 405–410. 10.1044/jshr.2803.405

Ortman JM, Velkoff VA, & Hogan H (2014). An aging nation: the older population in the United States (pp. 25–1140). Washington, DC: United States Census Bureau, Economics and Statistics

Administration, US Department of Commerce. Retrieved from: http://usd-apps.usd.edu/coglab/schieber/psyc423/pdf/AgingNation.pdf

Ostrand R, & Gunstad J (2020). Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults. Journal of Geriatric Psychiatry and Neurology. 10.1177/0891988720933358

Pakhomov SV, Eberly L, & Knopman D (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. Neuropsychologia, 89, 42–56. 10.1016/j.neuropsychologia.2016.05.031 [PubMed: 27245645]

Pistono A, Jucla M, Barbeau EJ, Saint-Aubert L, Lemesle B, Calvet B, … & Pariente J (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer’s disease. Journal of Alzheimer's disease, 50(3), 687–698. DOI: 10.3233/JAD-150408

Pistono A, Jucla M, Bézy C, Lemesle B, Le Men J, & Pariente J (2019). Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early Alzheimer's disease. International Journal of Language & Communication Disorders, 54(3), 390–400. 10.1111/1460-6984.12444 [PubMed: 30444044]

Pravatà E, Tavernier J, Parker R, Vavro H, Mintzer JE, & Spampinato MV (2016). The neural correlates of anomia in the conversion from mild cognitive impairment to Alzheimer’s disease. Neuroradiology, 58(1), 59–67. DOI 10.1007/s00234-015-1596-3 [PubMed: 26400852]

Reitan RM. (1958). Validity of the Trail Making Test as an Indicator of Organic Brain Damage. Perceptual and Motor Skills, 8(3):271–276. doi:10.2466/pms.1958.8.3.271

Rosselli M, & Ardila A (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. Brain and Cognition, 52(3), 326–333. 10.1016/S0278-2626(03)00170-2 [PubMed: 12907177]

Saffran EM, Berndt RS, & Schwartz MF (1989). The quantitative analysis of agrammatic production: Procedure and data. Brain and language, 37(3), 440–479. 10.1016/0093-934X(89)90030-8 [PubMed: 2804622]

Saykin AJ, Gur RC, Gur RE, Shtasel DL, Flannery KA, Mozley LH, … & Mozley PD (1995). Normative neuropsychological test performance: effects of age, education, gender and ethnicity. Applied Neuropsychology, 2(2), 79–88. 10.1207/s15324826an0202_5 [PubMed: 16318528]

Slegers A, Filiou R-P, Montembeault M, & Brambati SM (2018). Connected Speech Features from Picture Description in Alzheimer’s Disease: A Systematic Review. Journal of Alzheimer’s Disease, 65(2), 519–542. 10.3233/JAD-170881

Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, & Markesbery WR (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. Jama, 275(7), 528–532. doi:10.1001/jama.1996.03530310034029 [PubMed: 8606473]

Teng EL & Chui HC (1987). The Modified Mini-Mental State (3MS) examination. Journal of Clinical Psychiatry, 48(8), 314–318. [PubMed: 3611032]

Toth L, et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Current Alzheimer’s Research, 15, 130–138. 10.2174/1567205014666171121114930

Valcour V, Masaki K, Curb J, et al. (2000). The Detection of Dementia in the Primary Care Setting. Archives of Internal Medicine, 160, 2964–2968. doi:10.1001/archinte.160.19.2964 [PubMed: 11041904]

Weschler D (2008). The Wechsler Adult Intelligence Scale. San Antonio. TX: Pearson Assessments.

Williams BW, Mack W, & Henderson VW (1989). Boston Naming Test in Alzheimer’s disease. Neuropsychologia, 27(8):1073–1079. doi:10.1016/0028-3932(89)90186-3 [PubMed: 2797414]

Table 1.

Demographic, Medical, and Neuropsychological Characteristics of the Sample.

	Intact	MCI	Test Statistic	*p*	Cohen’s d
Demographic/Medical					
Age	67.81 ± 8.43	68.58 ± 6.58	t(86) = 0.42	.68	0.10
Gender (% Female)	68.9%	65.4%	$\chi^2(1) = 0.10$	.75	
Years of Education	15.47 ± 2.38	15.27 ± 3.00	t(86) = 0.33	.74	0.07
Hypertension	33.8%	53.8%	$\chi^2(1) = 3.05$	.08	
Type 2 Diabetes	11.3%	19.2%	$\chi^2(1) = 0.98$	.32	
Depression	16.1%	26.9%	$\chi^2(1) = 1.37$	.24	
Neuropsychological Testing					
3MS	58.19 ± 6.23	53.19 ± 11.64	t(86) = 2.62	.001	0.54
HVLT Total Learning	51.71 ± 7.96	44.00 ± 7.86	t(86) = 4.16	<.001	0.97
HVLT Delay	50.32 ± 10.69	41.04 ± 10.71	t(86) = 3.72	<.001	0.86
HVLT Discrimination	52.21 ± 7.66	42.31 ± 10.72	t(86) = 4.90	.001	1.06
CFT Immediate Recall	56.57 ± 13.77	40.04 ±15.98	t(86) = 4.84	<.001	1.15
CFT Delayed Recall	54.66 ± 13.17	37.76 ± 15.46	t(86) = 5.15	<.001	1.18
Digit Forward - Longest String	51.72 ± 9.49	49.12 ± 10.30	t(86) = 1.14	.001	0.26
Digit Backward – Longest String	55.00 ± 8.76	49.31 ± 10.39	t(86) = 2.62	.001	0.59
Trail Making Test A	54.95 ± 8.08	47.88 ±10.02	t(86) = 3.47	.001	0.78
Trail Making Test B	53.64 ±5.86	47.80 ± 8.57	t(86) = 3.64	<.001	0.80
Frontal Assessment Battery	54.65 ± 11.71	41.50 ±15.34	t(86) = 4.37	<.001	0.96
Controlled Oral Word Association Test	57.39 ± 11.64	55.54 ±9.21	t(86) = 0.72	.47	0.80
Animal Naming	57.89 ±12.21	48.65 ±9.60	t(86) = 3.43	.001	0.84
Boston Naming Test – Short Form	58.68 ±8.25	54.73 ± 10.25	t(86) = 1.90	.06	0.42

Table 2.

Sample-wise Characteristics for Each Speech Feature.

Speech Index	Intact	MCI	t (df = 86)	p	Cohen’s d
Total words	491.55 ± 233.54	348.23 ± 241.29	2.60	.01	0.60
Filler words	0.64 ± 0.61	0.73 ± 0.80	0.55	.58	0.13
Empty words	0.25 ± 0.16	0.18 ± 0.12	1.91	.06	0.49
Definite articles	1.63 ± 0.55	1.26 ± 0.64	2.78	.007	0.62
Indefinite articles	0.51 ± 0.17	0.47 ± 0.19	0.90	.37	0.22
Pronouns	2.54 ± 0.72	2.12 ± 0.77	2.47	.02	0.56
Nouns	4.43 ±1.11	3.51 ± 1.12	3.55	.001	0.83
Verbs	4.43 ± 1.14	3.81 ± 1.28	2.22	.03	0.51
Determiners	2.51 ± 0.73	1.98 ± 0.80	3.04	.003	0.69
Content Words	9.81 ± 2.49	8.06 ± 2.89	2.87	.005	0.64
Lexical Frequency	5.42 ± 0.33	5.62 ± 0.52	2.23	.03	0.46
Type-Token ratio	0.40 ± 0.07	0.44 ± 0.09	2.52	.01	0.50
Honoré’s statistic	−6.10 ± 1.44	−7.86 ± 3.30	3.50	.001	0.69
Brunet’s index	13.40 ± 0.98	12.70 ± 1.36	2.74	.008	0.59
Speech rate	2.41 ± 0.34	2.31 ± 0.43	1.11	.27	0.26
Filler rate	0.07 ± 0.06	0.09 ± 0.08	1.23	.22	0.28

Note. Values show mean and standard deviation of each linguistic feature for each participant group, as well as the t-statistic, p-value, and Cohen’s d for the comparison between the two groups.

Table 3.

Lexical-semantic features that were calculated on the transcripts of spontaneous speech produced by participants retelling the Cinderella story.

Feature name	Description
Total words	Overall count of all phonological entities spoken; including real words, nonwords, and partial words
Filler words	Count of filled pauses (e.g., “uh”, “um”, “hmm”), as a percentage of total word count
Empty words	Count of empty words (e.g., “thing”, “place”, “stuff”), as a percentage of total word count
Lexical frequency	Mean of the log of the frequency of all real words spoken
Type-token ratio	Ratio of unique words (types) to total words (tokens) spoken, used as a measure of vocabulary size and lexical diversity; higher values means the speaker produced a more varied vocabulary
Honoré’s statistic	Measure of lexical richness/diversity based on the number of words produced exactly once; higher values mean more diverse speech. It is calculated as: (100 * log(tokens)) / (1 - V_1/types), where V_1=number of words spoken exactly once
Brunet’s index	Measure of lexical richness (i.e., degree of variation in vocabulary), which is less biased by text length, calculated from the total number of words produced (tokens) and the number of unique words (types); lower values mean richer speech. It is calculated as: tokens ^ types ^ (−0.165)
Speech rate	Count of total words divided by total elapsed time of the speech (in words per second)
Filler rate	Count of filler words divided by total elapsed time of the speech (in words per second)
Definite articles	Count of uses of “the”, as a percentage of total word count
Indefinites articles	Count of uses of “a” and “an”, as a percentage of total word count
Pronouns	Count of pronouns, as a percentage of total word count
Nouns	Count of nouns, as a percentage of total word count
Verbs	Count of verbs, as a percentage of total word count
Determiners	Count of determiners, as a percentage of total word count
Content words	All words that are not function words (as defined by the list of stop words in NLTK), as a percentage of total word count

Table 4.

Multiple Logistic Regression Using All Speech Features Jointly to Predict MCI Status.

				95% Confidence Interval for Exp(β)	
	β	Wald	Exp(β)	Lower Bound	Upper Bound
Intercept	81.45	1.33			
Total words	0.00	0.00	1.00	0.98	1.02
Filler words	6.99	4.28	1089.78	1.45	821255.69
Empty words	−3.38	1.60	0.03	0.00	6.37
Definite articles	4.88	4.01	131.37	1.11	15572.28
Indefinite articles	3.43	1.95	30.92	0.25	3841.58
Pronouns	−0.01	0.00	0.99	0.13	7.28
Nouns	−2.70	4.30	0.07	0.01	0.86
Verbs	2.44	4.64	11.42	1.25	104.65
Determiners	−2.83	1.45	0.06	0.00	5.88
Content words	0.11	0.01	1.11	0.19	6.64
Lexical frequency	−0.78	0.41	0.46	0.04	5.05
Type-token ratio	−48.52	0.90	8.49 x 10^{-22}	3.00 x 10^{-65}	2.40 x 10^{+22}
Honoré’s statistic	−0.55	1.78	0.58	0.26	1.30
Brunet’s index	−5.17	1.81	0.01	3.03 x 10^{-06}	10.65
Speech rate	1.32	1.28	3.74	0.38	36.66
Filler rate	−45.83	2.53	1.25 x 10^{-20}	3.68 x 10^{-45}	42149.03

Table 5.

Simple Logistic Regression Using Each Speech Feature Individually to Predict MCI Status.

	Wald	Exp(β)	% Intact Correctly Predicted	% MCI Correctly Predicted	*p*
Total words	5.87	1.00	98.4	11.5	.015
Filler words	0.31	1.21	100.0	0.0	.58
Empty words	3.46	0.04	100.0	0.0	.06
Definite articles	6.74	0.32	95.2	15.4	.009
Indefinite articles	0.81	0.29	100.0	0.0	.37
Pronouns	5.46	0.44	98.4	11.5	.019
Nouns	9.84	0.46	91.9	19.2	.002 *
Verbs	4.52	0.63	98.4	11.5	.033
Determiners	7.78	0.36	95.2	23.1	.005 *
Content words	7.04	0.76	96.9	19.2	.008
Lexical frequency	4.40	3.53	96.8	15.4	.036
Type-token ratio	5.51	1314.58	95.2	15.4	.019
Honoré’s statistic	7.94	0.66	95.2	19.2	.005 *
Brunet’s index	6.38	0.68	95.2	19.2	.012
Speech rate	1.22	0.49	100.0	0.0	.27
Filler rate	1.43	49.03	98.4	3.8	.23

Note.

*indicates p<.05 when corrected for multiple comparisons using the Benjamini-Hochberg (1995) method.