AWARD NUMBER: W81XWH-20-1-0531

TITLE: Automated Speech Analysis in FTD Spectrum Disorders

PRINCIPAL INVESTIGATOR: Murray Grossman

CONTRACTING ORGANIZATION: University of Pennsylvania

REPORT DATE: August 2022

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release.
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
August 2022	Annual	1AUG2021 - 31JUL2022

4. TITLE AND SUBTITLE

Automated Speech Analysis in FTD Spectrum Disorders

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-20-1-0531

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Murray Grossman, Naomi Nevler

E-Mail: mgrossma@pennmedicine.upenn.edu, naomine@pennmedicine.upenn.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania
Philadelphia, PA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Frontotemporal degeneration (FTD) is an understudied form of focal dementia. Its public health impact is immense because clinical FTD is the most common neurodegenerative disease in individuals <65 years old. FTD presents with a specific language deficit (Primary Progressive Aphasia, PPA). A careful analysis of everyday speech can help identify variants of PPA. This proposal fills a major gap by providing an objective, replicable, fully automated approach to discerning speech characteristics of PPA. FTD may co-occur with a motor disorder, including Amyotrophic Lateral Sclerosis (ALS) and Chronic Traumatic Encephalopathy (CTE) which are directly relevant to the military. Detailed analyses of speech in FTD spectrum disorders with associated motor impairments are rare, and we propose to extend our analyses to FTD patients with motor disorders. Finally, longitudinal analyses of speech can play an important role in prognosis and in treatment trials, but longitudinal studies are rare. This study pursues these issues with three Specific Aims: 1. Develop an automated algorithm to analyze lexical semantic word-level content and grammatical category in FTD; 2. Develop automated algorithms to align lexical content with acoustic signal in connected speech samples of FTD speakers; and 3. Develop algorithms to automatically characterize the properties of the complex (acoustic and lexical) signals that are associated with sentence boundaries and syntactic units in FTD speech.

15. SUBJECT TERMS
Frontotemporal dementia, primary progressive aphasia, speech

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	48	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. Introduction

Frontotemporal degeneration (FTD) is an understudied form of focal dementia. Its public health impact is immense because clinical FTD is the most common neurodegenerative disease in individuals <65 years old. FTD presents with a specific language deficit (Primary Progressive Aphasia, PPA). A careful analysis of everyday speech can help identify variants of PPA. This proposal fills a major gap by providing an objective, replicable, fully automated approach to discerning speech characteristics of PPA. FTD may co-occur with a motor disorder, including Amyotrophic Lateral Sclerosis (ALS) and Chronic Traumatic Encephalopathy (CTE) which are directly relevant to the military. Detailed analyses of speech FTD spectrum disorders with associated motor impairments are rare, and we propose to extend our analyses to FTD patients with motor disorders. Finally, longitudinal analyses of speech can play an important role in prognosis and in treatment trials, but longitudinal studies are rare. This study pursues these issues with three Specific Aims: 1. Develop an automated algorithm to analyze lexical semantic word-level content and grammatical category in FTD; 2. Develop automated algorithms to align lexical content with acoustic signal in connected speech samples of FTD speakers; and 3. Develop algorithms to automatically characterize the properties of the complex (acoustic and lexical) signals that are associated with sentence boundaries and syntactic units in FTD speech.

2. Keywords

Frontotemporal dementia, primary progressive aphasia, semantic variant primary progressive aphasia, non-fluent/agrammatic primary progressive aphasia, behavioral variant frontotemporal dementia, speech, natural language processing

3. Accomplishments

Major Goals:

Major goals in the first year included: 1 - exploring, testing, and training automated part of speech (POS) tagging algorithms in FTD speech; 2 - testing automated dependency parsing in FTD speech; and 3 - reviewing and correcting aligned output of speech samples from untrained forced alignment (FA).
Major goals in the second year included: 1 – Validating FA performance in FTD speech samples; 2 – defining speech markers for boundaries of utterance and syntactic structures.

Accomplishments:

Our work was delayed in the first year due to delays in HRPO approval as well as ongoing limited campus activity due to the pandemic. Our study team was able to continue some remote work and most recently resumed activity on campus. Our center now works in a hybrid mode with continuous clinical data collection, supported in part by the digital infrastructure provided by our project.

1) Major activities:

(a) Tested and trained automated English part of speech (POS) tagger and dependency parser on FTD speech corpus (aim 1), (b) Tested Forced Aligner performance on FTD speech samples of picture description tasks (aim 2), (c) Explored lexical and acoustic markers of utterance boundaries (aim 3).

2) Specific objectives:

(a) Perform Exploratory Data Analysis (EDA) on untrained POS tags; (b) Characterize speech of different FTD phenotypes; (c) Validate characteristic speech features of FTD phenotypes with clinical measures; (d) use machine learning algorithms to train POS classifiers for FTD syndromes; (e) evaluate accuracy of automated POS tagger; (f) review and correct dependency parser analysis of FTD speech samples; (g) Review and correct aligned files from untrained forced alignment (FA); (h) Define the acoustic and lexical characteristics of utterance boundaries; (i) Define lexical acoustic markers of within-utterance syntactic unit boundaries (e.g. dependent clause).

3) Key outcomes:

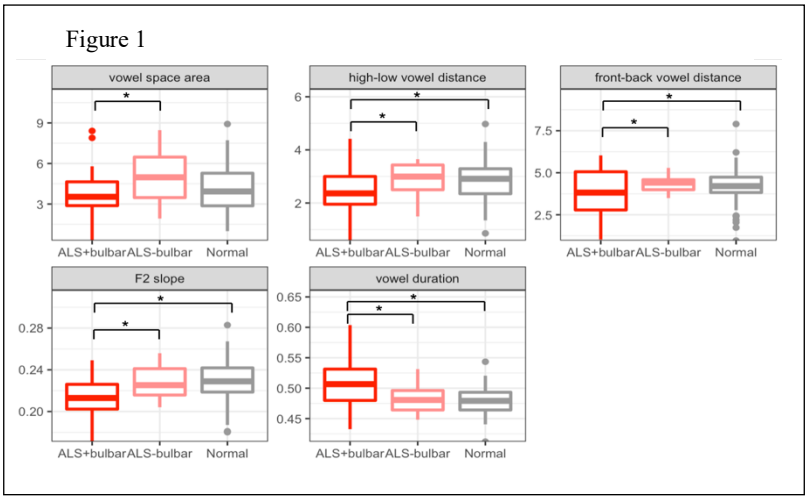
We defined well-characterized, distinct FTD speech patterns by FTD phenotype at the word level, reflecting words' lexical roles (Part-of-Speech, POS, e.g., nouns, verbs, adjectives) [see figure 1 in Cho et al. Cortex 2021]. We validated distinct speech features clinically, by linking them to impaired performance on neuropsychiatric tests and to atrophy in relevant areas of the brain cortex per structural MRI [see section 4.1 and figure 3 in Cho et al. Cortex 2021].

We successfully applied a Support Vector Machine (SVM) machine learning algorithm to train an automatic POS tagger on our FTD speech samples, resulting in speech classifiers for specific FTD phenotypes [see Cho et al. LREC 2020, figure 4].

We applied time series analysis methods to define the distinct longitudinal behaviors of speech patterns over time and disease progression in each FTD phenotype [see Nevler et al. Alzheimers Dement, 2020, figure 1].

We evaluated the performance of our current automatic Forced Aligner (FA). This advanced language tool is essential for our aim 3, where we align transcripts with the audio signal to extract acoustic features in relation to specific parts of the utterances that have syntactic meaning (e.g., dependent clause). In the past FA algorithms' outputs showed many instances of malalignments related to overlapping speakers. In our current evaluation we found no malalignments. With this FA output we extracted vowel specific frequency (formants) from our speech samples and identified some vowel related speech measures. These vowel speech measures capture the way we articulate by relating to the extent and speed of tongue movements in specific planes and speech of articulation. Impaired articulation is a hallmark of motor speech disorders, such as we see in Amyotrophic Lateral Sclerosis (ALS), where tongue and nasopharyngeal muscle weakness involving the vocal apparatus result in slurred and dysarthric speech. We identified specific vowel measure (figure 1) impairment in speakers with ALS compared with normal speakers and speakers with a pure behavioral syndrome (bvFTD). We pursued clinical validation by relating these vowel measures specifically to the existence and severity of bulbar disease in ALS (figure 2). We also related a composite score derived from these vowel measures to cortical atrophy in the motor tongue regions of the brain in the precentral gyrus using structural MRI scans available from most of these speakers with ALS (figure 3).

Figure 1



In aim 3 we manually annotated transcripts for syntactic phrase structures (sentence, dependent clause, etc.). This was done independently by two expert linguists, while discussing ambiguous cases until agreement was reached. We then explored acoustic features that relate to pausing and to expert labeled syntactic boundaries. We identified acoustic markers that relate

Figure 2

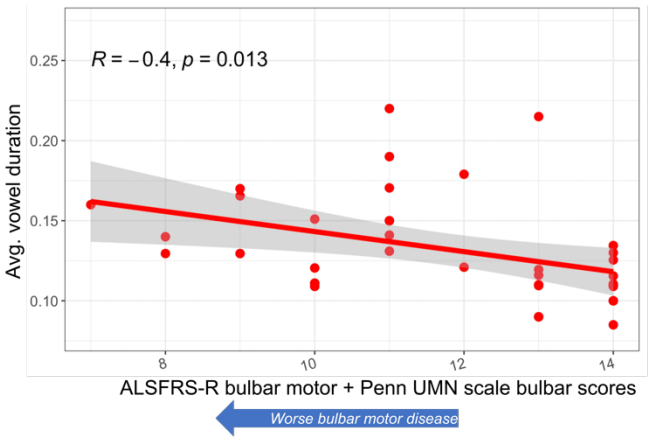
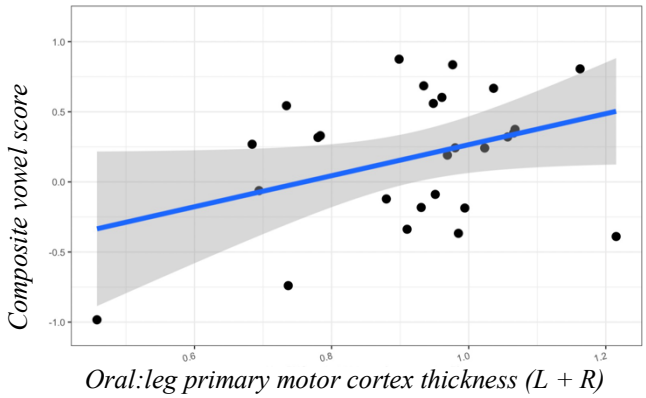


Figure 3



to syntactic boundaries. Specifically, words preceding such boundaries (e.g. clauses, conjunctions, sentence start) had longer duration and a rise in pitch contours, compared to words that did not precede a boundary. We also identified specific word POS that increase the likelihood of a following or preceding boundary. Combining these word-level lexical acoustic features (POS, word duration and pausing) resulted in an automated classifier of syntactic phrase boundary with an AUC of 0.94.

Based on our syntactic boundary analysis, we are currently pursuing an additional study, where we are investigating different approaches to automatically capturing the degree of grammatical complexity of utterances. Grammatical complexity is impaired in many neurodegenerative conditions, but especially in people with FTD syndromes, where one of the variants is presented predominantly with agrammatism. Agrammatism affects the ability to express oneself with complex sentences and to comprehend complex messages and instructions and this leads to extreme disability in daily living because of difficulty with communication. The speech in disorders such as Alzheimer's disease also becomes simplified grammatically and this could be an early sign for neurodegeneration. Expert training is currently required to properly define grammatical structures in speech and objective quantification is challenging. Thus, developing automated measures to detect and quantify simplified grammar is essential for implementing this early marker in clinical assessments. Automatically identifying syntactic structures (aim 3) is an essential first step in this endeavor.

Some of these findings have already been published and made available on the national medical library and we are currently working on three additional manuscripts to report the more recent findings.

4) Other achievements:

Since the beginning of this project and particularly in the past year we have expanded our speech data collection to include the picnic scene picture and other, more traditional neuropsychiatric tasks from each participant per visit. We find in our preliminary studies added value in recording and digitizing the responses of some neuropsychiatric tests. These include verbal fluency tasks, which inform us about semantic and phonetic linguistic processing as well as executive strategies, passage reading, which

informs on articulation, and story recall, which informs on memory functions. We also developed and validated our speech measures across recording environments, mainly, in-person (with the use of tablets and smartphones) and remotely (with virtual conference applications). We developed an in-house recording application for in-person recording, that ensures high quality recording and secure data transfer to our secured servers. This is a complete make-over of our digital speech data flow, supporting speech and neuropsychiatric data collection under any condition with minimal burden to patients, caregivers and examiners. We have generalized all specifications so that they can be customized to fit other centers and protocols. We have already experimented with these data flow pipelines and provided service to other centers in our department.

In another study using our word-level standardized lexical acoustic markers from the cookie theft picture descriptions, we compared speech in patients with biological evidence of Alzheimer's disease (per autopsy or CSF biofluid marker profile) between those who presented with aphasia (the logopenic variant of primary progressive aphasia, lvPPA, one of the phenotypes described in FTD) and those who presented with a typical amnestic syndrome. We found a distinct speech pattern of linguistic impairment in lvPPA compared with amnestic AD [Cho et al. Neurology 2022, figure 1], and we also identified a common speech pattern for both phenotypes [Cho et al. Neurology 2022, figure 2]. This common pattern relates to more general non-linguistic deficits and potentially to their common underlying AD pathology.

Opportunities for Training and Professional Development:

While there was no formal intent to provide opportunities for professional development, there are postdoctoral fellows who benefit greatly from learning about design and execution of multidisciplinary research projects. This includes regular weekly meetings, regular scientific presentations of project progress, and sharing these at scientific conferences (see listed presentations below). We also included invited guest speakers in our weekly meetings to learn about other research efforts in the field and to facilitate potential future collaborations. Our postdoctoral fellows gained new connections and learning opportunities with relevant colleagues at Penn and outside of it, including computational linguists, expert neurologists, and speech language pathologists.

Dissemination of Results:

In addition to peer-reviewed publications (listed below), our team members presented our work regularly within our department (Neurology), at other departments within the university of Pennsylvania as well as

outside of UPenn (Washington University, Miami University CReATe consortium annual meeting, University of California San Francisco). We also hold an outreach annual conference for caregivers of FTD patients, where we present our work to the general public.

Next Reporting Period Activities to Accomplish Goals:

Our major goals in the next year will include: (a) completing the investigation and clinical validation of FA vowel extraction system; (b) improving accuracy of automatic syntactic boundary detection system; (c) testing and validating automated measures of grammatical complexity in the speech of FTD patients; (d) submitting manuscripts on automated utterance boundary markers, automated grammatical complexity measures and vowel measures.

4. Impact

Impact on Principal Discipline:

There is increasing interest in developing digital biomarkers to serve as meaningful clinical outcome assessment tools in clinical trials for neurodegenerative conditions. We recently participated in the first inaugural Holloway summit for digital biomarkers, organized by the Association for FTD and took a leading role in organizing the meeting. Researchers were invited from different fields in Academia around the world, technology and pharma industries as well as patient advocates. During this meeting, it became apparent that digital speech measures are especially appreciated in the research community. Speech and language measures derived with advanced language technologies, such as the measures we are developing in the current project, are expected to become highly informative and useful in screening for clinical trials, monitoring response to treatment and aiding patient care in the home environment.

Impact on Other Disciplines:

Since the beginning of this project, we have engaged in many collaborations within Penn and outside of it that involve speech data. This includes a collaboration with the FTD center at UCSF, Alzheimer's disease (AD) research centers at Penn and Mt. Sinai medical center, Penn Parkinson's disease (PD) center. We share our standardized operating procedures for speech data collection and provide training and support to collaborators' teams. Processed speech data will be used in multiple observational studies in these centers and will facilitate research in the natural disease progression of these neurodegenerative conditions.

Impact on Technology Transfer:

There is increased understanding in the neurodegenerative research community of the value and importance of standardizing digital speech measures across centers, languages, and cultures to optimize cost-effective outcome measures for multi-center treatment trials for neurodegenerative conditions. Our work, emphasizing the development of objective, reproducible and standardized measures, will facilitate data sharing and technology transfer in the future, supporting the implementation of digital speech tools in clinical settings.

Impact on Society:

Because speech is so easily collected with minimal burden to the subject and can even be collected remotely, society will benefit from widespread standardized speech analysis methods to track cognitive decline and the development of neurodegeneration.

5. Changes/Problems

Changes in Approach:

Our evaluation of the current LDC in-house automated Forced Aligner (aim 2) showed higher accuracy than we originally expected when developing aim 2. The validated output seems adequate for the current project's needs and we were able to move forward with developing our articulatory pipeline, which relies on FA vowel extraction (FAVE). This objective goes beyond our original tasks in aim 2. We were also able to focus our effort on the study of utterance boundaries (aim 3), which is in much earlier stages of development.

Recent changes in the university's contracts required us to discontinue the use of BlueJeans, a video conference application and move to Zoom. This required adjustment of our remote speech data collection protocols. Though this was an unexpected change which required some investment on our part, we were able to make the required adjustments in time for the transition with no extra burden on our budget.

Problems or Delays and Actions or Plans for Resolution:

Our project was delayed by almost six months due to delays in HRPO approval, which was granted in March 2021. Additionally, the covid pandemic restricted activity on UPenn campus. This affected our ongoing clinical data collection, which is now continuing in a hybrid work mode. Most speech data is collected in-person and we are also collecting speech remotely. We are now close to pre-pandemic clinical visits capacity, and we maintain contingency plans for remote speech data collection as well. We validated the precision of our digital speech measures across these different conditions (in-person versus remote).

Changes Impacting Expenditures: Nothing to Report.

Changes in Human Subjects: Nothing to Report.

6. Products

Publications, conference papers, and presentations:

1. Cho S, Cousins KAQ, Shellikeri S, Ash S, Irwin DJ, Liberman MY, Grossman M, Nevler N. Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology. Neurology, Jul 2022, 99 (4) e313-e322.
2. Cho S, Shellikeri S, Ash S, Liberman MY, Grossman M, Nevler, N. Automatic classification of AD versus FTLD pathology using speech analysis in a biologically confirmed cohort. Alzheimer's Dement. 2021 17: e052270.
3. Cho, S., Agmon, G., Shellikeri, S., Cousins, K., Ash, S., Irwin, D., Spindler, M., Madiedo, A.D.A., Elman, L., Quinn, C., Liberman, M., Grossman, M., Nevler, N. (2022) Prosodic characteristics of prepausal words produced by patients with neurodegenerative disease. Proc. Speech Prosody 2022, 120-124.
4. Gonzalez-Recober C, Cho S, Liberman M, Grossman M and Naomi Nevler. (July 2022). Application of advanced language technologies in analysis of category naming fluency task in healthy participants. Poster presentation at the 2022 Alzheimer's Association International Conference.
5. Cho S, Shellikeri S, Ash S, Agmon G, Cousins KAQ, Gonzalez-Recober C, Nevler N, Liberman M, Grossman M. (July 2022) Longitudinal changes of automated speech markers in MCI and mild AD. Poster presentation at the 2022 Alzheimer's Association International Conference.
6. Cho S, Nevler N, Parjane N, Cieri C, Liberman M, Grossman M, Cousins K. (2021). Automated analysis of letter fluency data. Poster presented at the 2021 Society for Neurobiology of Language annual meeting.
7. Shellikeri S, Cho S, Liberman M, McMillan C, Elman E, Ash S, Grossman M, Nevler N. Longitudinal speech markers of motor and cognitive disease in ALS-FTD spectrum. Poster presented at the 2021 Society for Neurobiology of Language annual meeting.
8. Cho S, Nevler N, Parjane N, Cieri C, Liberman M, Grossman M, Cousins KAQ. Automated Analysis of Digitized Letter Fluency Data. Front Psychol. 2021 Jul 29; 12:654214.
9. Cho S, Nevler N, Ash S, Shellikeri S, Irwin DJ, Massimo L, et al. Automated analysis of lexical features in frontotemporal degeneration. Cortex. 2021; 137:215-31.

10. Cho, S., et al. (2020). Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features. LREC 2020 Language Resources and Evaluation Conference 11-16 May 2020.
11. Parjane N, Cho S, Ash S, Cousins KA, Shellikeri S, Liberman M, Shaw LM, Irwin DJ, Grossman M, Nevler N. Digital speech analysis in progressive supranuclear palsy and corticobasal syndromes. Journal of Alzheimer's Disease. 2021 Jan 1;82(1):33-45.
12. Nevler N, Ash S, Cho S, Shellikeri S, Parjane N, Irwin DJ, Liberman MY, Grossman M. A longitudinal study of automated analysis of acoustic speech markers in FTD and PPA: Biomarkers (non-neuroimaging)/Longitudinal change over time. Alzheimer's & Dementia. 2020 Dec;16: e045315.
13. Nevler N, Ash S, Cho S, Shellikeri S, Parjane N, Irwin DJ, Liberman MY, Grossman M. Automated semantic speech analysis in AD and lvPPA: Biomarkers (non-neuroimaging)/novel biomarkers. Alzheimer's & Dementia. 2020 Dec;16:e045300.
14. Shellikeri S, Cho S, Ash S, Parjane N, Elman L, McMillan CT, Grossman M, Nevler N. Longitudinal changes of automated speech measures in natural connected speech in ALS: Biomarkers (non-neuroimaging)/Longitudinal change over time. Alzheimer's & Dementia. 2020 Dec;16:e043028.
15. Nevler N, Ash S, McMillan C, Elman L, McCluskey L, Irwin DJ, Cho S, Liberman M, Grossman M. Automated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. Neurology 2020 Sept; 95(12): e1629-e1639.

Websites: Nothing to Report.

Technologies: Nothing to Report.

Inventions: Nothing to Report.

Other Products: Nothing to Report.

7. Participants & Other Collaborating Organizations

Participants:

NAME	PROJECT ROLE	PERSON-MONTHS WORKED	CONTRIBUTIONS	FUNDING SUPPORT
M. Grossman	PI	1	Scientific design and report	NIH
M. Liberman	Co-I	1	Scientific design and report	NSF
N. Nevler	Postdoc	3	Digitized acoustic analysis, scientific design and report	NIH

S. Cho	Postdoc	10	Digitized Lexical analysis	AA
S. Shellikeri	Postdoc	12	Digitized analysis of motor speech	AAN
G. Agmon	Postdoc	12	Digitized analysis of utterance syntax	
W. Xu	Res. Coord.	1	Research program coordination	NIH
C. Gonzalez	Res. Coord.	8	Research data collection and processing	
S. Ash	Res. Specialist	3	Transcription supervision, expert linguistic assessment of automated lexical tags	NIH
B. Nelson	Res. Specialist	1	Database design and maintenance	NIH
N. Ryant	Res. Specialist	1	Algorithm programming	NSF
S. Kulich	Res. Specialist	1	Algorithm programing	NSF
J. Zehr	Res.Specialist	1	Software development	NSF
J. Fiumara	Res. Specialist	1	Project integration	NSF

Change in Other Support: nothing to report.

Other organizations: Nothing to Report.

8. Special Reporting Requirements

N/A.

9. Appendices

None.

CORTEX 137 (2021) 215–231

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Automated analysis of lexical features in frontotemporal degeneration

Sunghye Cho [a,*], *Naomi Nevler* [b], *Sharon Ash* [b], *Sanjana Shellikeri* [b], *David J. Irwin* [b], *Lauren Massimo* [b], *Katya Rascovsky* [b], *Christopher Olm* [b,c], *Murray Grossman* [b] *and Mark Liberman* [a]

[a] *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA*
[b] *Department of Neurology and Penn Frontotemporal Degeneration Center, University of Pennsylvania, Philadelphia, PA, USA*
[c] *Department of Radiology and Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, PA, USA*

ARTICLE INFO

Article history:
Received 26 February 2020
Reviewed 4 June 2020
Revised 26 July 2020
Accepted 22 January 2021
Action editor Stefano Cappa
Published online 6 February 2021

Keywords:
Frontotemporal degeneration
Lexical measures
Primary progressive aphasia
Part-of-speech
Aphasia

ABSTRACT

We implemented an automated analysis of lexical aspects of semi-structured speech produced by healthy elderly controls (n = 37) and three patient groups with frontotemporal degeneration (FTD): behavioral variant FTD (n = 74), semantic variant primary progressive aphasia (svPPA, n = 42), and nonfluent/agrammatic PPA (naPPA, n = 22). Based on previous findings, we hypothesized that the three patient groups and controls would differ in the counts of part-of-speech (POS) categories and several lexical measures. With a natural language processing program, we automatically tagged POS categories of all words produced during a picture description task. We further counted the number of *wh*-words, and we rated nouns for abstractness, ambiguity, frequency, familiarity, and age of acquisition. We also computed the cross-entropy estimation, where low cross-entropy indicates high predictability, and lexical diversity for each description. We validated a subset of the POS data that were automatically tagged with the Google Universal POS scheme using gold-standard POS data tagged by a linguist, and we found that the POS categories from our automated methods were more than 90% accurate. For svPPA patients, we found fewer unique nouns than in naPPA and more pronouns and *wh*-words than in the other groups. We also found high abstractness, ambiguity, frequency, and familiarity for nouns and the lowest cross-entropy estimation among all groups. These measures were associated with cortical thinning in the left temporal lobe. In naPPA patients, we found increased speech errors and partial words compared to controls, and these impairments were associated with cortical thinning in the left middle frontal gyrus. bvFTD patients' adjective production was decreased compared to controls and was correlated with their apathy scores. Their adjective production was associated with cortical thinning in the dorsolateral frontal and

* *Corresponding author*. Linguistic Data Consortium, 3600 Market Street, Suite 810, University of Pennsylvania, Philadelphia, PA, 19104-2653, USA.
E-mail address: csunghye@ldc.upenn.edu (S. Cho).
https://doi.org/10.1016/j.cortex.2021.01.012

orbitofrontal gyri. Our results demonstrate distinct language profiles in subgroups of FTD patients and validate our automated method of analyzing FTD patients' speech.

1. Introduction

Speech production is a complex, intentional, planned activity. Speakers select appropriate words from their lexicon that are consistent with the meaning of an intended message, arrange words in a specific order following the syntactic rules of the language, plan their articulations, and articulate the prepared message following the phonological rules of the language. This involves multiple brain regions, and we can expect patients with degenerative brain conditions to show impaired speech compared to healthy adults. Moreover, depending on the form of disease, we can expect distinct impairment profiles. In this study, we investigate linguistic impairments in patients with frontotemporal degeneration (FTD) by implementing a fully automated method of lexical analysis.

FTD refers to a group of disorders caused by atrophy in the brain's frontal, temporal, and parietal lobes, which is related to the underlying accumulation of abnormal Tau or TDP proteins. The disorders we investigated include two forms of primary progressive aphasia (PPA), the semantic variant PPA (svPPA) and the nonfluent/agrammatic variant PPA (naPPA). We also examined behavioral variant frontotemporal dementia (bvFTD). Patients with svPPA, also known as semantic dementia, are characterized by semantic impairment and difficulties in confrontation naming and lexical retrieval (Amici et al., 2007; Hodges & Patterson, 2007; Wilson et al., 2010). Previous studies have shown that svPPA patients have difficulty processing words denoting concrete objects (Bonner et al., 2009; Bonner, Price, Peelle, & Grossman, 2016; Breedin, Saffran, & Coslett, 1994; Cousins, Ash, Irwin, & Grossman, 2017; Cousins, York, Bauer, & Grossman, 2016; Macoir, 2009), but their prosody and syntax are less disrupted (Adlam, Bozeat, Arnold, Watson, & Hodges, 2006; Ash et al. 2006, 2009; Nevler, Ash, Irwin, Liberman, & Grossman, 2019; Thompson & Mack, 2014). It has also been observed that svPPA patients' lexical retrieval is related to word familiarity and frequency (Bird, Lambon Ralph, Patterson, & Hodges, 2000; Hodges & Patterson, 2007; Rogers, Patterson, Jefferies, & Lambon Ralph, 2015). Patients with naPPA, also known as progressive non-fluent aphasia, present with effortful speech, slow speech rate, grammatical simplification, and speech errors or apraxia of speech (AoS) (Ash et al., 2009; Grossman, 2012; Grossman et al., 1996; Josephs et al., 2006; Ogar, Dronkers, Brambati, Miller, & Gorno-Tempini, 2007). These patients may also have difficulty retrieving verbs (Hillis, Oh, & Ken, 2004; Hillis, Tuffiash, & Caramazza, 2002; Rhee, Antiquena, & Grossman, 2001). Patients with bvFTD undergo changes in personality and social cognition and also present impairments in behavior, such as apathy and disinhibition. Previous studies have reported that bvFTD patients have subtle linguistic deficits with reduced retrieval of abstract words, reduced speech rate, tangential speech with irrelevant subject matter, and limited narrative expression (Ash et al., 2006; Cousins et al., 2017; Farag et al., 2010; Gunawardena et al., 2010; Hardy et al., 2016).

While valuable, most previous studies have relied on subjective, manual assessments of speech, which require a substantial amount of time, labor, and cost. There are also potential difficulties with manually coding the part of speech (POS) categories of every token due to the time, effort, and expertise that are required, so previous studies involving POS analysis have rarely examined every word of an utterance. This is a problem in studying language use in patients with dementia, because many previous studies have shown that such patients tend to produce fewer words than controls (e.g., Ash et al., 2013; Slegers, Filiou, Montembeault, & Brambati, 2018; Tappen, Williams, Barry, & DiSesa, 2002). However, previous studies have failed to show in detail which POS categories were reduced in which patient groups due to the effort required for manual POS tagging. As a result, large-scale studies have rarely been performed. The present study describes implementation of a novel, quantitative, reproducible, automated approach to studying lexical characteristics of patients with FTD. We show that our novel methods are reliable with validation against manual gold-standard data. We also provide novel findings by directly examining all POS categories from a semi-structured speech sample elicited during a picture description task. Few studies have compared FTD subgroups on a variety of lexical measures and studied POS production in bvFTD; this is the first comprehensive assessment of POS expression in bvFTD of which we aware. We further focus on lexical characteristics of FTD patients' speech because the lexicon is important in verbal communication where the goal is to convey meaningful messages to interlocutors. We also examine two global text measures: cross-entropy and lexical diversity. Cross-entropy is a useful measure in understanding how predictable a text sample is, in comparison to much larger language samples, and lexical diversity represents the diversity in a speaker's vocabulary usage. Our novel, automated technique for text analysis is based on a modern natural language processing (NLP) program and examines speech samples in a large cohort of FTD patients.

Based on previous findings, we hypothesize that frequencies of POS categories as determined by an automated POS tagger and lexical measures are valuable in distinguishing the svPPA, naPPA, and bvFTD patient groups, as follows.

1) In svPPA, we expect that patients would produce fewer nouns but more pronouns than the other patients related to their impairment in confrontation naming. We also expect these patients to produce more *wh*-words (e.g., "What is this?"), since they have difficulty retrieving the names of objects or understanding a pictured object. We

also expect that their nouns would be different on some lexical measures from those produced by the other patient groups due to their semantic impairment. Also, because their speech includes more pronouns and abstract, ambiguous nouns, we expect the cross-entropy measure to be low, indicating more predictability. Furthermore, we expect these language characteristics to be related to regions of cortical thinning in the temporal lobe (e.g., Cousins et al., 2017, 2018; Wilson et al., 2010).

2) We expect that naPPA patients would differ from the other patient groups in their frequency of speech errors, partial words, due to AoS and their difficulty in retrieving verbs. We also expect these measures to be related to cortical thinning in the left frontal lobe (e.g., Ash et al., 2010).
3) In bvFTD, we expect to find reduced production of abstract words compared to the other groups. We also expect that bvFTD patients who are apathetic would not modify or elaborate on the details of objects, so bvFTD patients' use of fewer adjectives was expected to be related to level of apathy. Adverb counts might also be lower in apathetic bvFTD patients, but to a lesser degree than adjective counts, since adverbs do not always serve the same modifying and elaborating role that adjectives do. Also, we expect these measures will be related to cortical thinning in the frontal lobe (e.g., Massimo et al., 2015).
4) We expect all patients to differ from controls in lexical diversity, consistent with previous studies, which have often showed significantly decreased lexical diversity in brain-damaged patients compared to controls (e.g., Kavé & Dassa, 2018).

2. Methods

We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

2.1. Participants

We examined 138 patients with FTD diagnosed by experienced neurologists (M.G., D.J.I.) in the Department of Neurology at the Hospital of the University of Pennsylvania according to published criteria (Gorno-Tempini et al., 2011; Rascovsky et al., 2011). This includes 42 patients with svPPA, 22 patients with naPPA, and 74 patients with bvFTD. Among the svPPA patients, we included 32 cases with concomitant mild behavioral features, a common co-occurrence. These patients did not differ significantly from the other 10 svPPA patients without behavioral impairment in terms of demographic characteristics or linguistic performance. We also included 37 healthy seniors as a control group. The Institutional Review Board of the Hospital of the University of Pennsylvania approved the study, and written consent was obtained from all participants. The conditions of our ethics approval do not permit public archiving any raw data associated with this study. Readers seeking access to the data should contact Penn Frontotemporal Degeneration Center or one of the authors, Naomi Nevler. Access will be granted to qualified researchers in accordance with ethical procedures governing the reuse of sensitive data. Specifically, requestors must complete a formal data sharing agreement and regulatory approvals to obtain the data.

All participants (n = 175) were native speakers of English. The participants were matched on education level, but not on age and sex ratio (Table 1). A Tukey's post-hoc test of the ANOVA analysis revealed that bvFTD patients were significantly younger than naPPA patients and controls (*vs* naPPA, $p = .002$; *vs* control, $p = .007$). svPPA patients were also significantly younger (*vs* naPPA, $p = .007$; *vs* control, $p = .029$). Separate chi-squared tests indicated that there were more females in the control group than in the bvFTD group ($p = .006$) although the sex ratio was not different among the patient groups. One-way ANOVA tests showed that patient groups were matched on disease duration [$F(2,135) = 1.5$, $p = .24$] and Mini Mental State Exam [MMSE; $F(2,123) = .759$, $p = .47$].

We also measured patients' performance on neuropsychological assessments (Table 1) with the Boston Naming Test (BNT, Kaplan, Goodglass, & Weintraub, 2001), Pyramids and Palm Trees Test (PPT, Howard & Patterson, 1992), Animals and Tools Category Naming Fluency (Lezak, Howieson, & Loring, 1983) to assess semantic knowledge, and the Philadelphia Brief Assessment of Cognition (PBAC, Libon et al., 2011) to assess the degree of apathy in participants. Legal copyright restrictions prevent public archiving of the various instruments and test batteries used in this study, which can be obtained from the copyright holders in the cited references. As expected, on the BNT, in which participants are asked to name an object, svPPA patients had significantly lower scores than the other groups ($p < .001$ for all three pairwise comparisons). Patients with bvFTD also scored significantly lower on the BNT than healthy controls ($p = .01$). On PPT, where participants were asked to choose one of two words that was more closely related in meaning to a target word, svPPA patients had lower scores than controls ($p < .001$) and naPPA patients ($p = .012$), and bvFTD patients also scored lower than controls ($p < .001$). All patient groups performed poorly on the category fluency tasks, where participants were asked to name items in a given category (either animals or tools), compared to controls ($p < .001$ for all three pairwise comparisons). The difference in the fluency task scores between bvFTD and svPPA patients was also significant ($p < .001$). On the PBAC apathy scale, where the degree of apathy is assessed by interviewing family members or observing patients' behavior during the clinical interview (0 = most apathetic, 4 = least apathetic), the result of an ANOVA analysis was significant [$F(3,115) = 2.88$, $p = .039$], but pair-wise group comparisons were not significant. We further compared the number of participants who were apathetic (PBAC apathy score ≤ 2) and non-apathetic (PBAC apathy > 2) by group with chi-squared tests, and we found that there were more apathetic patients in bvFTD than in svPPA ($\chi = 6.09$, $p = .014$) and in the control group ($\chi = 6.46$, $p = .011$), but not compared to naPPA ($\chi = 2.44$, $p = .12$). The participants' demographic and neuropsychological characteristics are summarized in Table 1.

2.2. Picture description procedure

The participants were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination

Table 1 – Group means (SD) and omnibus test results of clinical and demographic characteristics. ANOVA analyses were used to compare all measures between groups except sex ratio, where a chi-squared test was used. MRI: Magnetic resonance imaging, BNT: Boston Naming Test, PPT: Pyramids and Palm Trees Test, PBAC: The Philadelphia Brief Assessment of Cognition (0 = most apathetic, 4 = least apathetic). Numbers in square brackets are Ns when less than the total.

	control (N = 37)	bvFTD (N = 74)	naPPA (N = 22)	svPPA (N = 42)	Group comparisons
Sex					
Female (N, percent)	24 (64.9%)	26 (35.1%)	11 (50%)	23 (54.8%)	$\chi = 9.9$, $p = .019$
Male (N, percent)	13 (35.1%)	48 (64.9%)	11 (50%)	19 (45.2%)	
Education	15.9 (2.5)	15.8 (2.8)	15.3 (3.1)	15.1 (2.8)	$F(3,171) = .9$, $p = .437$
Age (years)	68.5 (7.9)	63.1 (8.7)	70.4 (9.4)	63.3 (7)	$F(3,171) = 7.3$, $p < .001$
Disease duration (years)	–	4.4 (3.5)	3.2 (1.9)	3.9 (2)	$F(2,135) = 1.5$, $p = .239$
Time between MRI & picture	–	[42]	[8]	[26]	$F(2,73) = 1.1$, $p = .326$
description recording (months)	–	2.2 (1.9)	1.7 (1.7)	2.8 (2.6)	
Mini mental state exam (0–30)	[31]	[68]	[20]	[38]	$F(3,153) = 12.1$, $p < .001$
	29.2 (1)	23.6 (5.5)	22.7 (6)	22.1 (6.3)	
BNT (0–30)	[23]	[68]	[16]	[40]	$F(3,143) = 99.8$, $p < .001$
	27.9 (2.5)	23.8 (5.8)	24.7 (4.6)	7.5 (6.4)	
Animals and Tools (Max 60 sec)	[23]	[65]	[16]	[39]	$F(3,139) = 30.8$, $p < .001$
	16.8 (4.6)	9.2 (5.2)	8.2 (4.4)	5.1 (3.8)	
PPT (0–52)	[18]	[35]	[7]	[19]	$F(3,75) = 11.4$, $p < .001$
	50.8 (1.9)	42.9 (7.9)	48.4 (2.9)	39.6 (6.6)	
PBAC Apathy (0–4): N	[6]	[62]	[14]	[37]	$F(3,115) = 2.88$, $p = .039$
	3.3 (.5)	2.1 (1.1)	2.7 (1.2)	2.5 (1.2)	

(Goodglass & Kaplan, 1983), and the descriptions were digitally recorded. Patients were prompted to continue describing the picture, if necessary, following a silence of several seconds, and they were encouraged to continue up to about 60 sec after the beginning of the description. Recordings were orthographically transcribed by a linguist (S.A.), blinded to the clinical features and group membership of the participants, and further reformatted and time-stamped by trained, blinded annotators at the Linguistic Data Consortium (LDC) of the University of Pennsylvania. We note that no part of the study procedures or analyses were pre-registered prior to the research being conducted.

2.3. *POS tagging*

We employed spaCy (Honnibal & Johnson, 2015; https://spacy.io), an NLP library in Python, to automate the POS tagging process. spaCy has two different schemes of POS tagging. One is the OntoNotes 5 (Weischedel et al., 2013) version of the Penn Treebank tag set (Marcus, Santorini, & Marcinkiewicz, 1993). The other is the Google Universal POS tag set (Petrov, Das, & McDonald, 2012), which is simpler than the Penn Treebank scheme. The two POS tag schemes are not independent of each other, since spaCy maps the Penn Treebank tag to the simpler Google Universal POS tag set. Here we report the Universal POS tag results except for the calculation of the number of tense-inflected verbs, for which we used the Penn Treebank tags, because tense-inflected verbs are not distinguished by the broader Universal POS categories. The POS lists are included in the Appendix (Table A).

We wrote a Python program (S.C.) by which spaCy automatically tokenized each utterance in the transcripts with its default language model and annotated the POS category and the lemma for each word. In total, we had 21,990 tokenized words with both the Universal and Penn Treebank tags. The token count of each POS category (both Universal and Penn Treebank schemes) was tallied for each participant, and the number of each POS category per 100 words was calculated. We used POS counts per 100 words in all statistical analyses.

The Universal POS annotation scheme of spaCy uses "X" to tag words that do not exist in its language model. For example, *sptrkljgl* would be tagged as X, since the token is not a valid English word. Patients did not produce many non-English words during the picture description task, but they produced many partial words and speech errors, which looked like non-English words in the transcription. For example, in the utterance, "There's a *pu*-um a plate," ***pu***-was tagged as X by spaCy, since this is not an English word. We compared the frequency of this category by group in order to evaluate the frequency of speech errors and partial words in naPPA patients compared with other groups.

We also calculated the number of tense-inflected verbs per 100 words, the number of unique nouns per 100 words, the number of *wh*-words per 100 words and the total number of words in each speech sample, using the Penn Treebank POS tags and lemma counts. First, we summed all tokens produced by each participant for the total number of words. This measure included partial words and speech errors. The number of tense-inflected verbs was calculated by summing the number of modal auxiliary verbs, the number of past tense verbs, and the number of present tense verbs, using the Penn Treebank POS tags (Appendix Table A). This sum was used to compute the number of tense-inflected verbs per 100 words. We counted the number of unique lemmas in each speech sample and calculated the number of unique nouns per 100 words. We also counted the number of *wh*-words, "what" and "who", using a Python script, and calculated the number of *wh*-words per 100 words to examine the clinical observations that svPPA patients use more *wh*-words to ask objects' names than the other groups do, due to their impairments in object

knowledge. To see if the ratio of POS categories differed by group, we calculated the ratio of content words to function words for each participant. The calculated measures were used for between-group comparisons, covarying for age and sex.

2.4. Lexical measures

We performed additional analyses of nouns because of their potential value in distinguishing FTD patient groups. We rated nouns for abstractness on a continuum from concrete to abstract (Brysbaert, Warriner, & Kuperman, 2014), semantic ambiguity (number of a given word's meanings in a context, Hoffman, Lambon Ralph, & Rogers, 2013), word frequency (defined as word frequency per million words on a $\log_{10}$ scale, Brysbaert & New, 2009), age of acquisition (AoA) (Brysbaert, Mandera, & Keuleers, 2018) and word familiarity (z-standardized measure of the number of people who know a given word, Brysbaert et al., 2018). We wrote a Python program (S.C.) to provide these parameters automatically for all nouns that spaCy annotated. We built a pipeline in the program which (1) rated a word if it was listed in the published database and (2) rated the lemma of a word if the word was not listed in the published database but its lemma was (e.g., overflowed ⇒ overflow). The program excluded a word if neither the word nor its lemma was included in the lists (e.g., *countertop*, *Mary Jane*). This excluded about 3% of the words tagged as nouns (141 out of 4,157 words) from the analysis. The abstractness ratings ranged from 1 to 5, where the most concrete was 5 and the most abstract was 1. For clearer representation, we inverted the scale so that the most concrete was 1 and the most abstract was 5.

Along with these measures, we also computed cross-entropy estimation using all the words of the participants' picture descriptions. Cross-entropy estimation is a measurement that estimates the predictability of all words of a document with respect to their predictability in a larger language sample. High cross-entropy (uncertainty) is observed in a document that uses unusual words given the source language sample. A computational linguist (M.L.) computed the cross-entropy estimation of the speech samples by patients, based on a 1-g language model of three large-scale corpora: the SUBTLEXus (Brysbaert & New, 2009), Fisher English Training Speech (Cieri, Graff, Kimball, Miller, & Walker, 2004), and Switchboard (Godfrey & Holliman, 1997).

We also calculated lexical diversity for each patient. Traditionally, lexical diversity has been measured using the type/token ratio, where *type* is the number of unique words and *token* is the number of instances of each word. However, the type/token ratio has the disadvantage that the measure is affected by the total number of words. To address this problem, various approaches have been suggested by previous studies (e.g., Covington & McFall, 2010; Jarvis, 2002; McKee, Malvern, & Richards, 2000; Moscoso del Prado Martín, 2017; Tweedie & Baayen, 1998). In this study, we used the moving-average type/token ratio (Covington & McFall, 2010), which has been reported to be a stable measure for lexical diversity (Cunningham & Haley, 2020). It calculates a type/token ratio for a fixed-length window, moving one word at a time from the beginning to the end of a text document, and averages type/token ratios from all windows. We varied the length of the window from 20 to 35 words by 5-word increments. Since the results were the same regardless of the window size, we reported results from 20-word windows in Fig. 2 and Table 3. Hereafter, abstractness, ambiguity, frequency, familiarity, AoA, cross-entropy, and lexical diversity are referred to as "lexical measures". "Language measures" is used to refer to both POS counts and the lexical measures.

2.5. Imaging methods

High resolution T1 volumetric brain MRI data that were collected on a Siemens 3.0 T Trio scanner at 1 mm isotropic resolution were available for a subset of patients (n = 94): 18 controls, 42 bvFTD, 8 naPPA, and 26 svPPA patients. The mean time interval between MRI and speech sample collection was 1.95 months (SD = 2.11 months). Clinical and demographic characteristics of this subset of patients matched those of the patients in the full dataset, and the groups in this subset were matched on demographic characteristics. The demographic and language measurements of these patient groups are summarized in the Appendix (Tables B, C).

Sixty-five images were collected in an axial plane with repetition time = 1620 msec, echo time = 3.87 msec, slice thickness = 1.0 mm, flip angle = 15°, matrix = 192 × 256, and in-plane resolution = .9766 × .9766 mm. Twenty-nine images were collected with a sagittal acquisition with repetition time = 2300 msec, echo time = 2.95 msec, slice thickness = 1.2 mm, flip angle = 9°, matrix = 256 × 240, and in-plane resolution = 1.05 × 1.05 mm. Briefly, whole-brain MRI volumes were preprocessed using the antsCorticalThickness.sh processing pipeline, implemented using the Advanced Normalization Tools (ANTs) (https://github.com/ANTsX/ANTs; Tustison et al., 2014). Cortical thickness was estimated at each voxel of the cortex using the DiReCT algorithm (Das, Avants, Grossman, & Gee, 2009). easy_lausanne (https://github.com/mattcieslak/easy_lausanne; Daducci et al., 2012) run on our local template, which was created based on data from the Open Access Series of Imaging Studies (OASIS) (Marcus, Fotenos, Csernansky, Morris, & Buckner, 2007), to create a standard cortical parcellation. The template parcellation was then spatially normalized to each participant's native T1 space using the template-to-native T1 warps generated by ANTs, and then we calculated the mean cortical thickness in each region of interest (ROI) of the Lausanne 250 scale, which we used for our analysis.

To identify regions of atrophy in patients, we compared cortical thickness of all patients in each patient group with those of the controls for all cortical regions of interest (ROIs) and selected our specific ROIs for each patient group, where patients' cortical thickness was significantly thinner than that of the controls ($p < .01$ for svPPA and bvFTD, and $p < .05$ for naPPA, both uncorrected p-values). We applied a more lenient p-value threshold ($p < .05$) in selecting ROIs for naPPA patients due to the small number of patients with MRI data. We further identified ROIs that were significantly correlated with the degree of apathy (PBAC, Table 1) for bvFTD patients among the selected ROIs ($p < .05$) to mask the regressions. This method enabled us to restrict our interpretation of the regression results of adjectives in bvFTD to those brain regions that were

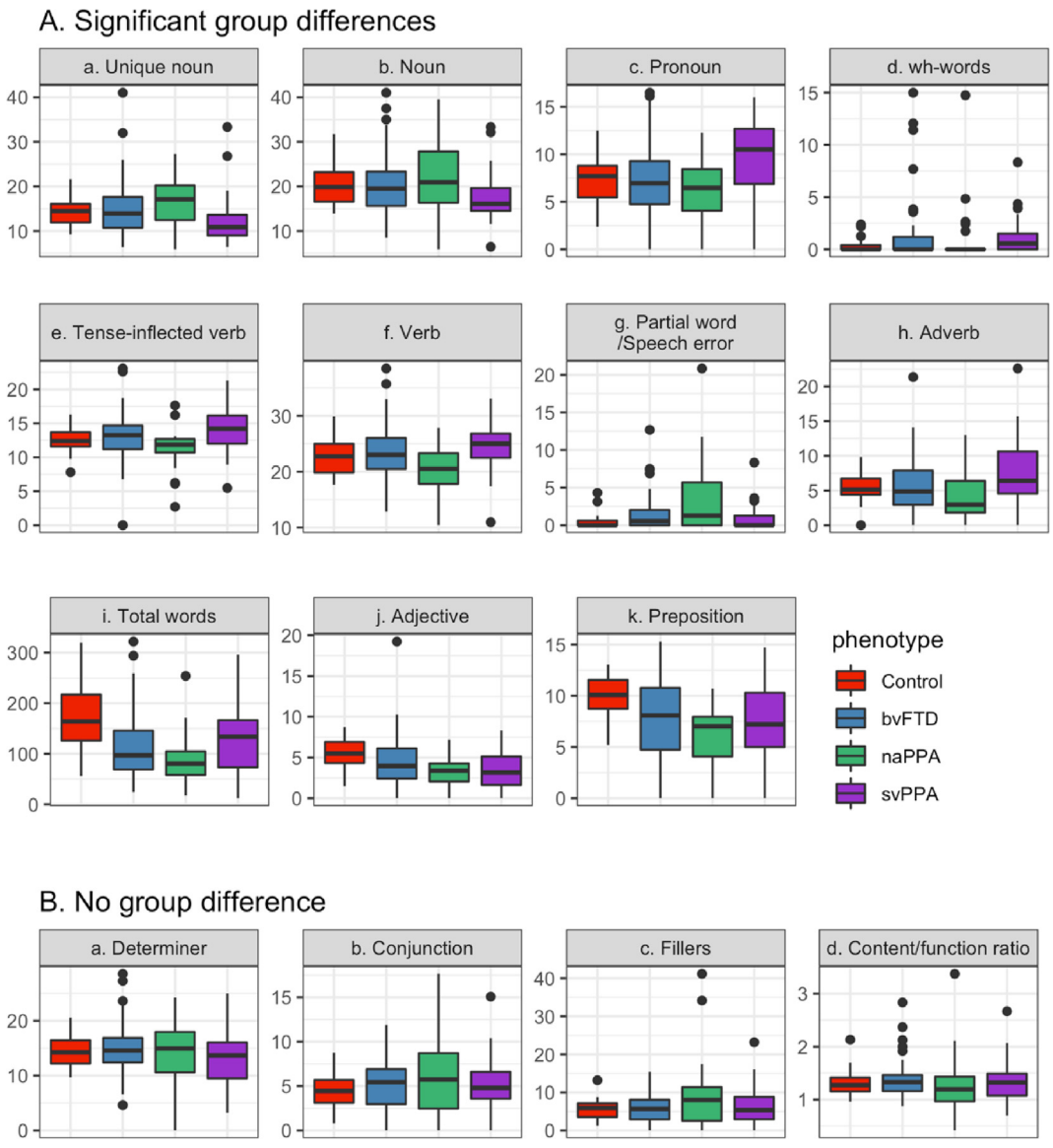


Fig. 1 – Median, 1 SD, 25th-75th percentile and outliers in POS categories per 100 words, total number of words and the ratio of content words by phenotype.

significantly related to apathy. Adverbs were not considered in the MRI analyses since the apathy scores were not significantly correlated with adverb production for bvFTD patients.

2.6. *Statistical considerations*

Since the abstractness, ambiguity, frequency, familiarity, and AoA measures were rated for each noun, we averaged those values per individual and used mean per measure per participant in the ANCOVAs. We did not average cross-entropy and lexical diversity measures, since these were global measures (only one value per individual). Levene's test for homogeneity of variance, residuals, and Q–Q plots were employed to validate the requirements for parametric tests. Group comparisons were performed with Analysis of Covariance (ANCOVA) for the frequency of each POS category per 100 words and each of the lexical measures as a dependent variable, with phenotype as an independent variable. We introduced age and sex as covariates in the group comparison analyses of all language measures, as the groups were not matched on these factors. For those measures where the requirements for parametric tests were not met, we performed the rank-based inverse normal transformation (Conover, 1980) on the values of language measures, and the transformed values were used as the dependent variable in an ANCOVA. When there was a significant group effect, pair-wise group comparisons were conducted with the *lsmeans* package (Lenth, 2016) in R to adjust for multiple comparisons with false discovery rate. Since the group difference from ANCOVA was marginal in the counts of nouns and adverbs per 100 words, we performed logistic regressions as supplementary analyses with age and sex as covariates to compare the number of patients who had a z-score < -1 by group, where the z-score scale was computed based on the controls' mean and standard

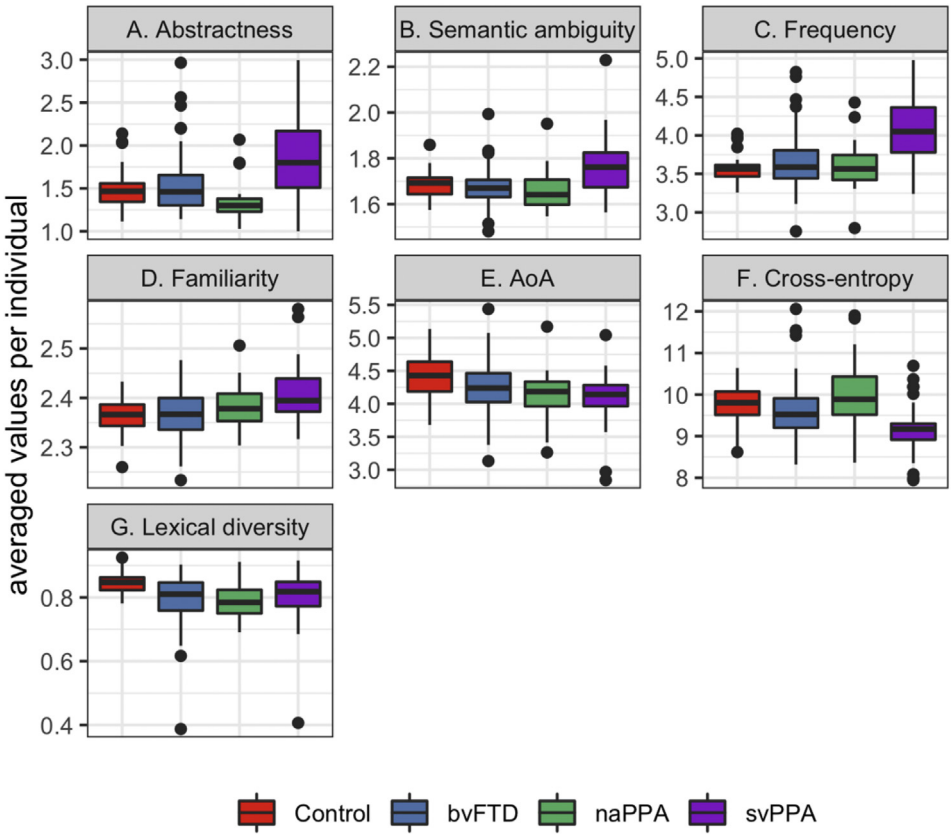


Fig. 2 – Median, 1 SD, 25th-75th percentile and outliers of abstractness scores, semantic ambiguity ratings, word frequency, word familiarity, and age of acquisition of nouns; and cross-entropy estimation and lexical diversity across all words.

Table 2 – Group means (SD) and omnibus test results from ANCOVA analyses of the POS categories per 100 words, total number of words, and the ratio of content words of all participants.

		Control	bvFTD	naPPA	svPPA	F	*p*
Significant group differences	Unique nouns	14.7 (3.19)	14.87 (5.93)	16.73 (5.96)	12.21 (5.19)	$F(3,169) = 3.46$	.018
	Nouns	20.32 (4.4)	20.16 (6.48)	21.92 (8.7)	17.49 (5.3)	$F(3,169) = 2.52$	.058
	Pronouns	7.33 (2.41)	7.13 (3.77)	6.46 (3.2)	9.74 (3.9)	$F(3,169) = 7.66$	<.001
	wh-words	.34 (.53)	.6 (1.12)	.34 (.99)	1.61 (1.72)	$F(3,169) = 9.26$	<.001
	Tense-inflected verbs	12.47 (1.83)	12.94 (3.68)	11.26 (3.2)	14.14 (2.98)	$F(3,169) = 3.92$	.01
	Verbs	22.56 (3.42)	23.59 (4.86)	20.22 (4.42)	24.44 (4.06)	$F(3,169) = 3.86$	.011
	Speech errors/partial words	.48 (.89)	1.42 (2.26)	3.67 (3.4)	.89 (1.54)	$F(3,169) = 4.18$	.007
	Adverbs	5.59 (2.07)	6.04 (4.36)	4.37 (3.61)	7.05 (3.36)	$F(3,169) = 2.82$	.041
	Total words	174.38 (66.38)	109.99 (62.35)	91 (55.8)	127.57 (66.5)	$F(3,169) = 11.37$	<.001
	Adjectives	5.54 (1.82)	3.98 (3.16)	3.17 (2.03)	3.69 (2.04)	$F(3,169) = 5.87$	<.001
	Prepositions	9.96 (1.94)	7.63 (4.06)	5.98 (3.19)	7.24 (3.72)	$F(3,169) = 7.66$	<.001
No group differences	Determiners	14.16 (2.48)	14.85 (4.33)	14.34 (5.4)	13.35 (4.98)	$F(3,169) = .97$	.41
	Conjunctions	4.43 (1.91)	5.12 (2.69)	5.9 (4.68)	4.85 (2.88)	$F(3,169) = 1.41$	.24
	Fillers	5.5 (2.56)	5.89 (3.9)	10.03 (10.3)	6.27 (4.83)	$F(3,169) = 1.46$	.23
	Ratio of content to function words	1.32 (.22)	1.36 (.33)	1.3 (.6)	1.32 (.36)	$F(3,169) = .7$	.55

deviation. For the supplementary analysis for noun counts, we coded participants who produced fewer nouns (z-score < −1) as 1 and others as 0 for a dependent variable and ran a logistic regression with svPPA patients as a reference group and phenotype as an independent variable, controlling for age and sex. For the supplementary analysis of adverb counts, we coded participants who produced fewer adverbs as 1 and others as 0, with the naPPA group as our reference. We selected these reference groups based on the ANCOVA results. A Pearson's correlation test was performed to relate adjective and adverb counts to the apathy scores on the PBAC for each group to test our hypothesis for bvFTD patients. A series of separate linear regression analyses were performed to relate the cross-entropy estimations to each of the five lexical measures: abstractness, ambiguity, frequency, familiarity, and AoA.

Table 3 – Group means (SD) and omnibus test results from ANCOVA analyses of the lexical measures. AoA: Age of acquisition.

	Control	bvFTD	naPPA	svPPA	F	*p*
Abstractness (noun)	1.52 (.76)	1.55 (.83)	1.4 (.59)	1.92 (1.14)	F(3,169) = 11.68	<.001
Ambiguity (noun)	1.65 (.25)	1.64 (.26)	1.64 (.23)	1.74 (.28)	F(3,169) = 11.01	<.001
Frequency (noun)	3.39 (.86)	3.52 (.91)	3.44 (.91)	3.94 (.95)	F(3,169) = 12.99	<.001
Familiarity (noun)	2.38 (.14)	2.38 (.16)	2.39 (.14)	2.4 (.16)	F(3,169) = 3.81	.011
AoA (noun)	4.51 (1.42)	4.36 (1.33)	4.21 (1.24)	4.15 (1.13)	F(3,169) = 4.27	.005
Cross-entropy	9.72 (.49)	9.61 (.66)	9.9 (.84)	9.1 (.79)	F(3,169) = 7.7	<.001
Lexical diversity	.85 (.03)	.79 (.09)	.79 (.06)	.81 (.09)	F(3,169) = 6.21	<.001

Linear regression analyses were also used to relate the language measures to cortical thinning. We implemented univariate multiple regression analyses, covarying for potential confounding factors: the pulse sequence type used for MRI acquisition, patients' age, and disease duration. We did not covary for sex because the participants with MRI data did not significantly differ in the sex ratio across groups and there was no consistent evidence of the effect of sex on cortical thinning. The regions selected for svPPA and naPPA patients were used to relate their regions of cortical thinning to language measures that significantly differed between groups. The regions that were significantly related to the apathy scores in bvFTD patients were used to relate cortical thinning to adjective counts per 100 words. We report t-statistics at a significance level of .05 (two-tailed, uncorrected) for these regressions. All statistical analyses were performed in R (R Core Team, 2019) version 3.5.2 and RStudio (RStudio Team, 2016) version 1.1.456 (S.C.).

3. Accuracy validation of spaCy POS tags

Despite the fact that the accuracy of POS tagging reported by spaCy is very high (about 97%; https://spacy.io/models/en), it was not clear how well it would perform for a clinical dataset with abnormal speech. The training data (OntoNotes 5; Weischedel et al., 2013) of spaCy included natural conversations, but the ratio of conversational speech to written texts was only around 8.3% (120 K out of 1.4 million words) and the conversations were between healthy adults. To validate the accuracy of the spaCy POS tags on natural speech of a clinical population with abnormal speech, a linguist (S.A.) who was blinded to the automated analysis manually tagged a random subset of the transcripts comprehensively using the Google Universal POS scheme (6 Controls, 5 naPPA, 7 svPPA, and 7 bvFTD; 25 cases in total; 14.3% of the full dataset) to generate a gold standard dataset. We compared the results of spaCy in the same POS scheme to our gold standard dataset to calculate the error rates.

The error rate was generally low in all groups. The overall accuracy of spaCy on this subset of the picture description data was 91.1%, and the variances between the groups were not significantly different [Levene's test for homogeneity of variance: $F(3,21) = 2.69$, $p = .072$]. Also, a one-way ANOVA test revealed that the difference in error rates between the groups was not significant [$F(3,21) = 2.695$, $p = .075$]. The mean error rate of the control group was 5.4% (SD = 1.7%). The error rates of individual svPPA, naPPA, and bvFTD patients were slightly higher than that of the controls [svPPA: 8.8% (SD = 2.8%); naPPA: 13.3% (SD = 9.2%); bvFTD: 9.0% (SD = 3.0%)], but the difference among the patients groups was not significant [$F(2,16) = 1.32$, $p = .3$]. While the error rates for svPPA and bvFTD did not differ from that of controls, the difference between naPPA patients and the controls was significant ($p = .049$). This was expected, since naPPA speech contains the largest number of speech errors and partial words (see below) and thus differs most from the training data of spaCy.

For further validation, we correlated the token counts of nouns, tense-inflected verbs, and speech errors/partial words from spaCy with the counts that a linguist (S.A.) manually coded for all 175 participants. For the correlation between the noun counts of each individual, we used all NOUN tokens in the Universal tag set. Modal auxiliaries (MD), past (VBD) and present (VBP, VBZ) tense verbs in the Penn Treebank tag set were used for the correlation with tense-inflected verb counts. For speech errors, we compared the X category in the Universal tag set with the counts of manually coded speech errors. We found that the noun and inflected verb counts of spaCy and counts of those categories in our manual coding were strongly correlated (nouns: $r = .958$, $p < .001$; verbs: $r = .973$, $p < .001$). Also, the correlation of counts of X with our manual coding of speech errors was significant ($r = .43$, $p < .001$), suggesting that the POS tags produced by spaCy were reliable.

4. Results

We first present the results of automatic POS tagging (Section 4.1). Next, we show the group differences in the lexical measures (Section 4.2). In Section 4.3, we present the regression results with MRI data.

4.1. *POS categories and derived measures*

Table 2 summarizes the ANCOVAs comparing the POS measures per 100 words across the four groups. The groups differed significantly in the number of unique nouns (Fig. 1Aa). svPPA patients produced fewer unique nouns than naPPA patients ($p = .022$) and marginally fewer than bvFTD patients ($p = .056$). Noun production marginally varied by phenotype after controlling for age and sex (Fig. 1Ab). However, group-wise paired comparisons failed to reach significance (svPPA *vs* bvFTD: $p = .062$; svPPA *vs* naPPA: $p = .062$). A supplementary analysis with a logistic regression revealed that there were significantly more svPPA patients who produced fewer nouns (z-score < −1) compared to bvFTD patients ($z = -2.01$, $p = .044$)

and controls ($z = -2.75$, $p = .006$) but not compared to naPPA patients ($z = -1.67$, $p = .096$). Pronoun production (Fig. 1Ac) significantly differed between groups; pronouns were more frequent for svPPA patients than for the other groups (svPPA *vs* control: $p = .016$; svPPA *vs* naPPA: $p = .005$; svPPA *vs* bvFTD: $p = .002$). Also, the groups differed in the number of *wh*-words per 100 words (Fig. 1Ad). Patients with svPPA produced more *wh*-words than the other groups ($p < .001$ for all three pairwise comparisons).

The number of tense-inflected verbs per 100 words differed significantly by group (Fig. 1Ae). Pairwise group comparisons revealed that naPPA patients produced fewer tense-inflected verbs than svPPA patients ($p = .006$). Similarly, the group difference in the total number of verbs was significant (Fig. 1Af). naPPA patients produced fewer verbs than svPPA patients ($p = .008$) and bvFTD patients ($p = .016$). The groups were also different in the counts of speech errors and partial words (Fig. 1Ag). naPPA patients produced this category significantly more frequently than controls (naPPA *vs* control: $p = .005$). Adverb production also differed by group (Fig. 1Ah). naPPA patients tended to produce fewer adverbs than svPPA patients ($p = .052$). A supplementary analysis with logistic regression showed that the number of naPPA patients who produced fewer adverbs (z-score < -1) was greater than the number of svPPA patients ($z = -3.05$, $p = .002$) and controls ($z = -3.57$, $p < .001$) but not greater than the number bvFTD patients ($z = -1.8$, $p = .07$). The adverb counts per 100 words were not significantly correlated with apathy scores in any of the four groups ($p > .05$).

The total number of words participants produced during the picture description differed significantly by group (Fig. 1Ai). Controls produced significantly more words than any of the patient groups (*vs* bvFTDi: $p < .001$, *vs* naPPA: $p < .001$, *vs* svPPA: $p = .006$). Similarly, adjective production per 100 words significantly varied by group (Fig. 1Aj), and all patient groups used fewer adjectives than controls (*vs* bvFTD: $p = .013$; *vs* naPPA: $p = .003$; *vs* svPPA: $p = .002$). Furthermore, bvFTD patients' adjective counts per 100 words were significantly correlated with their apathy scores ($r = .32$, $p = .01$). The correlations of adjective production and apathy scores were not significant in the other three groups, and bvFTD patients' apathy scores were not significantly correlated with the other POS categories. The group difference in prepositions (Fig. 1Ak) was significant. Each patient group produced fewer prepositions than controls (*vs* bvFTD: $p = .004$; *vs* naPPA: $p < .001$; *vs* svPPA: $p = .004$). The differences among the patient groups for these categories were not significant.

The productions of conjunctions, determiners, fillers and the ratio of content to function words did not differ by group (Fig. 1B).

4.2. *Lexical measures*

All participants produced nouns that were not abstract in the picture description task, which is not surprising given the task of describing a picture that contains concrete objects. Yet, the group differences in abstractness were significant (Fig. 2A). svPPA patients produced nouns that were more abstract (i.e., less concrete) compared to bvFTD patients ($p < .001$), naPPA patients ($p < .001$), and controls ($p = .001$).

Semantic ambiguity ratings of nouns also differed significantly by group (Fig. 2B). Nouns produced by svPPA patients showed higher semantic ambiguity than those produced by all other groups (*vs* bvFTD: $p < .001$; *vs* naPPA: $p < .001$, *vs* controls: $p = .008$).

Patients tended to use more frequent nouns than controls, and the group difference in the frequency of nouns was highly significant (Fig. 2C). svPPA patients produced more frequent nouns than bvFTD patients, naPPA patients, and controls ($p < .001$ for all three pairwise comparisons).

The familiarity of nouns also significantly differed by group (Fig. 2D). svPPA patients used more familiar nouns than bvFTD patients ($p = .02$).

All patients tended to produce nouns acquired at an earlier age than controls (Fig. 2E), and the group difference in the age of acquisition of nouns was significant. svPPA patients produced nouns that were acquired earlier than controls ($p = .007$).

The cross-entropy estimation differed significantly by phenotype (Fig. 2F); the cross-entropy estimation of svPPA patients was lower than that of bvFTD patients ($p = .006$), naPPA patients ($p < .001$), and controls ($p = .001$). In other words, words produced by svPPA patients were more predictable than those produced by the other groups. To further examine why svPPA patients' cross-entropy estimation was lower than those of the other groups, five separate linear regression analyses were performed to relate cross-entropy estimation in svPPA patients to abstractness, ambiguity, frequency, familiarity, and AoA of nouns they produced. We found that abstractness, ambiguity, and word frequency were significantly related to cross-entropy estimation in svPPA (abstractness: $\beta = -.63$, $p < .001$, word frequency: $\beta = -.88$, $p < .001$; semantic ambiguity: $\beta = -2.8$, $p = .019$).

There was a significant group difference in lexical diversity that was measured by the moving-average type/token ratio with a window size of 20 words (Fig. 2G). Elderly controls showed higher lexical diversity than all patient groups (*vs* bvFTD: $p < .001$, *vs* naPPA: $p = .019$, *vs* svPPA: $p = .019$). When we tried different window sizes (25 words and 30 words), we found the same group differences (25-word window: *vs* bvFTD: $p = .002$, *vs* naPPA: $p = .004$, *vs* svPPA: $p = .018$; 30-word window: *vs* bvFTD: $p = .001$, *vs* naPPA: $p = .006$, *vs* svPPA: $p = .019$).

4.3. *MRI results*

Since patients showed significant differences from each other on the language measures, we examined the relations between cortical thinning and specific language measures in each group. We found distributions of cortical thinning that were representative of each group (Ash et al., 2009, 2012; Cousins et al., 2016; Massimo et al., 2009). The MRI results showed that svPPA patients had significant cortical thinning in the anterior temporal and orbital frontal cortex areas of both hemispheres, but cortical thinning was more prominent in the left hemisphere than the right hemisphere ($p < .01$; Fig. 3A). naPPA patients had significant cortical thinning most prominently in the left middle frontal, inferior temporal and middle temporal regions, but also apparent in the left supramarginal gyrus, right temporal gyrus, and right pars opercularis ($p < .05$, Fig. 3B). bvFTD patients had significant cortical thinning in the frontal

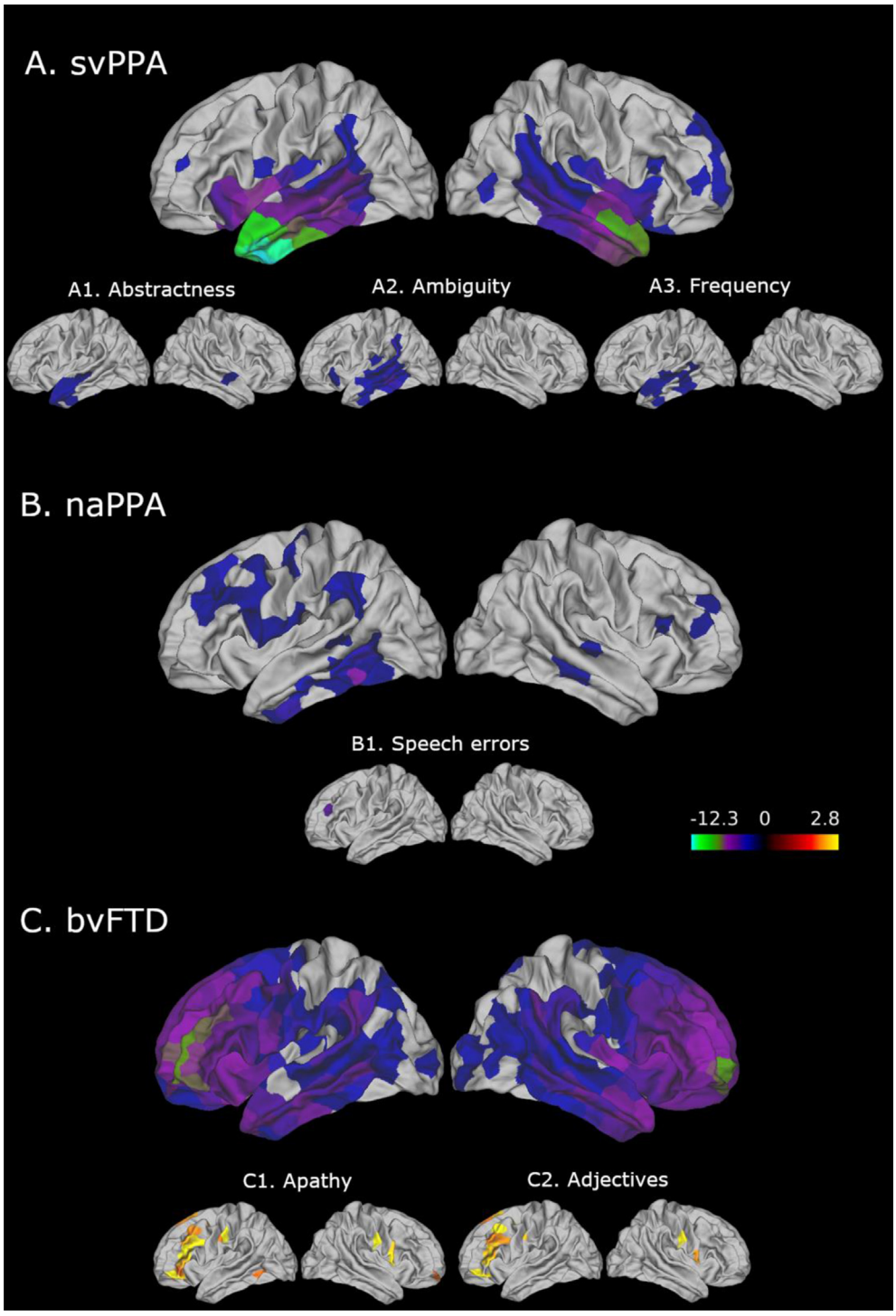


Fig. 3 – Cortical thinning in svPPA (A), naPPA (B) and bvFTD (C) patients, and areas with cortical thinning that were significantly related to linguistic measures ($p < .05$, uncorrected) in svPPA (A1-3), naPPA (B1), and bvFTD (C2) patients. Please note that these images are for illustration, and the complete results are summarized in Table 4.

and temporal lobes of both hemispheres ($p < .01$; Fig. 3C). We examined patients' speech production in relation to cortical thinning in greater detail, as summarized in Table 4. Examples of the associations are illustrated in Fig. 3.

We selected the language measures that were distinctive of svPPA patients in our main analyses outlined above. These showed significant associations with cortical thinning in anterior and middle temporal regions of the left hemisphere (Table 4). Fig. 3 shows brain images for the regions of cortical thinning associated with abstractness, ambiguity, and frequency that are frequently described for svPPA in the literature (Fig. 3A1-3).

We also found that the production of speech errors and partial words was related to cortical thinning in the left rostral

Table 4 – Results of regression analyses with cortical thinning in patients. L: left, R: right.

svPPA	Estimate	Std. Error	t-value	*p*-value
Noun				
L inferior temporal	.059	.021	2.85	.01
L middle temporal	.054	.026	2.09	.049
L superior temporal	.045	.021	2.18	.041
L insula	.035	.016	2.19	.04
Pronoun				
L inferior temporal	−.098	.038	−2.54	.019
L parahippocampal	−.059	.028	−2.16	.043
L entorhinal	−.104	.05	−2.08	.049
Wh-words				
L inferior temporal	−.219	.08	−2.6	.021
L middle temporal	−.244	.11	−2.14	.044
L superior temporal	−.19	.087	−2.2	.039
L fusiform	−.303	.108	−2.818	.01
L insula	−.142	.065	−2.207	.04
Abstractness				
L temporal pole	−.582	.228	−2.55	.019
L inferior temporal	−.531	.218	−2.42	.025
L middle temporal	−.652	.225	−2.89	.011
L superior temporal	−.51	.189	−2.69	.019
L fusiform	−.597	.243	−2.49	.027
R superior temporal	−.309	.14	−2.21	.038
Semantic ambiguity				
L inferior temporal	−2.609	.833	−3.11	.007
L middle temporal	−2.617	.896	−2.96	.011
L bank superior temporal	−1.795	.572	−3.13	.006
L superior temporal	−1.946	.693	−2.8	.013
L supramarginal	−1.722	.601	−2.86	.018
L insula	−.5	.205	−2.39	.026
L lateral orbitofrontal	−1.182	.564	−2.1	.048
Word frequency				
L inferior temporal	−.627	.258	−2.46	.024
L middle temporal	−.685	.264	−2.58	.019
L bank superior temporal	−.379	.176	−2.16	.043
L superior temporal	−.49	.208	−2.34	.031
L fusiform	−.593	.267	−2.22	.037
Word familiarity				
L inferior temporal	−.755	.29	−2.61	.016
L middle temporal	−.83	.247	−3.41	.009
L superior temporal	−.53	.182	−2.98	.018
L rostral middle frontal	−.821	.216	−3.8	.001
R rostral middle frontal	−.608	.222	−2.72	.014
L precentral	−.599	.163	−3.67	.001
L supramarginal	−.517	.19	−2.72	.013
L lateral orbitofrontal	−.365	.163	−2.24	.001
R superior frontal	−.592	.218	−2.72	.013
R pars opercularis	−.549	.192	−2.86	.009
Cross-entropy estimation				
L inferior temporal	.451	.187	2.4	.027
L middle temporal	.419	.199	2.1	.048
L bank superior temporal	.348	.143	2.45	.026
L superior temporal	.392	.156	2.51	.02
L fusiform	.713	.224	3.18	.004
naPPA	Estimate	Std. Error	t-value	*p*-value
Speech errors/Partial words				
L rostral middle frontal	−.194	.044	−4.39	.022
bvFTD	Estimate	Std. Error	t-value	*p*-value
Adjectives				
L orbitofrontal	.07	.028	2.56	.015
L rostral middle frontal	.05	.024	2.28	.031
L superior frontal	.07	.03	2.29	.03
L caudal middle frontal	.07	.025	2.67	.035
L post central	.06	.023	2.49	.018
R pre central	.07	.033	2.22	.032
R post central	.07	.025	2.74	.009

(*continued on next page*)

middle frontal gyrus for naPPA patients (Fig. 3B1), suggesting that speech errors and partial words are related to impairment in frontal executive functions. We also related verb, tense-inflected verb, and adverb counts to cortical thinning in naPPA patients, but the results were not significant.

The areas that showed a significant relation of cortical thinning to apathy in bvFTD patients (Fig. 3C1) are also significantly and positively related to their adjective production (Fig. 3C2), but no other POS categories. These areas include the left rostral and caudal middle frontal, the left superior frontal, and orbitofrontal regions.

5. Discussion

In this study, we examined word production and lexical measures of speech in FTD patients with a novel, automated method that is objective, comprehensive and reproducible. The POS counts derived from the Universal tag set were highly correlated with manually coded POS tags (Section 3). Moreover, distinct language measures were associated with each patient group (Sections 4.1–4.2). We found that svPPA patients produced fewer unique nouns than naPPA patients, and these nouns were more ambiguous, abstract, and frequent than those of naPPA and bvFTD patients. Correspondingly, svPPA patients produced more pronouns and *wh*-words. A new measure of cross-entropy estimation showed that their word selection in general was more predictable from its context than that of the other groups, and this was likely associated with noun abstractness, ambiguity, and frequency. Patients' words were less diverse than those of controls, but there was no significant group difference among the patient groups. naPPA patients produced fewer adverbs and more speech errors and partial words than the other groups. bvFTD patients produced fewer adjectives than controls, and their adjective production was significantly correlated with apathy scores. We also found significant associations between our language measures and cortical thinning. Cortical thinning in left anterior inferior and middle temporal gyri was associated with language measures in svPPA, and cortical thinning in the left middle frontal gyrus was associated with speech errors and partial words in naPPA. Cortical thinning in the left dorsolateral frontal and orbitofrontal gyri was associated with decreased adjective production in bvFTD. We discuss these findings in turn below.

5.1. *Lexical characteristics in svPPA*

The profiles of svPPA patients' nouns exhibited characteristics that significantly differed from those of the other groups. They

displayed high abstractness, semantic ambiguity, word frequency, and word familiarity. This is in line with other findings consistent with the hypothesis attributing the deficit in svPPA in part to the degradation of visual feature knowledge associated with object concepts (Bird et al., 2000; Bonner et al., 2009, 2016; Cousins, Ash, Olm, & Grossman, 2018; Cousins et al., 2016, 2017; Hoffman et al., 2013), which is due to cortical thinning in the left anterior and inferior temporal regions of the brain. This region constitutes a portion of visual association cortex which may contribute to the representation of visual feature knowledge associated with object concepts. It may explain in part why svPPA patients produced nouns with high abstractness in our results: abstract nouns are less dependent on visual feature knowledge to derive their meaning, thereby reducing the need to activate the anterior and inferior temporal regions of the brain. We also found that an increase in the abstractness rating of nouns was related to cortical thinning in the left anterior temporal region. In the context of concrete noun difficulty due to degraded representations of visual objects, it is not surprising that svPPA patients may substitute more pronouns, and this was reflected in associations with cortical thinning in the left temporal lobe and pronoun usage.

Previous observations have showed that svPPA patients' lexical retrieval is strongly graded by word familiarity and frequency (Bird et al., 2000; Hodges & Patterson, 2007; Rogers et al., 2015). These observations suggest that at least some proportion of the svPPA patients' picture description deficit is due in part to a lexical retrieval deficit that extends beyond their degraded semantic representations of object knowledge. As for semantic ambiguity, Hoffman et al. (2013) argue that this feature is highly correlated with abstractness ratings ($|r| = .51$, $p < .001$; Hoffman et al., 2013), suggesting that abstract words, such as *set* or *time*, are more ambiguous than concrete words, such as *desk* or *orange*. Given the high correlation of ambiguity and abstractness, it is not surprising that svPPA patients produced more nouns that were abstract and ambiguous. It is also possible that svPPA patients produce nouns such as *furniture, object,* or *thing* that are superordinate in a hierarchically organized semantic network because they do not have access to more concrete words. These possibilities need to be studied in future work.

Previous work describing the hub-and-spoke model (Patterson, Nestor, & Rogers, 2007) claims that disease in the anterior temporal lobe is responsible for a universal semantic deficit in svPPA. We found in the present study that svPPA patients used verbs more frequently than patients with naPPA. A frequent use of a specific POS category does not necessarily reflect the integrity of the meaning of this word class. However, on the assumption that patients use words with which they are more familiar in a semistructured speech sample, the more frequent use of verbs than nouns in svPPA would be contrary to the claim that the meaning of all words is degraded in svPPA. Likewise, we have showed that the meaning of words for abstract nouns is relatively preserved in svPPA (Bonner et al., 2016; Cousins et al., 2016) and that the meaning of words dependent on number knowledge is relatively preserved in svPPA (Ash et al., 2016). In a longitudinal study of lexical expression in svPPA, we found progressively reduced use of concrete words relative to abstract words (Cousins et al., 2018). Findings such as these are more consistent with a relatively selective degradation of the lexicon in svPPA. Additional work is needed to assess these claims.

5.2. *Lexical characteristics in naPPA*

A distinguishing feature of naPPA is that these patients produced more speech errors and partial words than other groups did. The increased speech error and partial word rate in naPPA conforms to previous findings that naPPA patients exhibit effortful and non-fluent speech (Ash et al., 2009, 2013; Croot, Ballard, Leyton, & Hodges, 2012; Gorno-Tempini et al., 2004; Grossman et al., 1996; Weintraub, Rubin, & Mesulam, 1990). We related increased partial words and speech errors to cortical thinning in the left middle frontal gyrus, which is in line with previous findings (Ash et al., 2009; Gorno-Tempini et al., 2004; Grossman et al., 1996). An important characteristic of naPPA patients is their AoS, that is, the poor coordination of the motor articulators during speech production (Ash et al., 2009; Gorno-Tempini et al., 2011; Grossman et al, 1996, 2005; Josephs et al., 2006; Ogar et al., 2007). It is claimed that a subset of naPPA patients has AoS without grammatical impairments, and that this differs from naPPA patients with grammatical impairments who have AoS (Josephs et al, 2012, 2013). A major challenge to this area of investigation is the ability to detect speech errors in an objective, reliable and reproducible manner. A rating scale based on subjective judgments has been developed, but reliability is challenging (Josephs et al., 2012; Strand, Duffy, Clark, & Josephs, 2014). Another challenge is that partial words in naPPA patients are not explained solely by AoS. Additional work is needed to confirm the identification of speech errors and partial words in an naPPA cohort, to extend this observation to patients with movement disorders such as progressive supranuclear palsy and corticobasal syndrome, and to distinguish this from speech errors in patients with bulbar disease such as amyotrophic lateral sclerosis.

Patients with naPPA in our study produced fewer verbs than the other groups. Decreased verb use in naPPA patients has also been observed in previous studies (Ash et al, 2009, 2013). Several accounts have been forwarded to explain this finding. One suggestion is that naPPA patients have difficulty producing tense-inflected verbs and constructing complex sentence structures due to a syntactic deficit, which leads to a reduced use of verbs in their speech (Ash et al., 2009, 2013; Grossman et al., 1996; Grossman et al., 2005). Alternatively, disease in naPPA may also affect motor association regions of the frontal lobe and interfere with the representation of action knowledge associated with verbs of action (Hillis et al, 2002, 2004). Yet another possibility is that the entire class of verbs is associated with a richer and more demanding set of features—including not only its semantic attributes but also a rich set of grammatical and thematic properties—and naPPA

patients have limitations in executive functioning that may make verbs more difficult for naPPA patients to process (Kramer et al., 2003; Libon et al., 2007; Weintraub et al., 1990). Previous work based on a smaller cohort of patients has suggested that the latter explanations are less likely than the grammatical one (Gunawardena et al., 2010), and we could not provide further evidence on these competing claims since the verb counts were not associated with cortical thinning in naPPA patients in our results. Additional work is needed to assess these claims.

5.3. *Lexical characteristics in bvFTD*

We hypothesized that bvFTD patients would differ in the counts of adjectives due to apathy and also that their nouns would be less abstract than those in the other groups. Our results showed that bvFTD patients produced fewer adjectives compared to controls, and their decreased adjective production was significantly correlated with their apathy scores, suggesting that bvFTD patients with fewer adjectives tended to be more apathetic. We identified regions of cortical thinning that were significantly related to apathy, including the left dorsolateral frontal and orbitofrontal gyri, and this result is in line with previous studies (Massimo et al., 2009, 2015). Furthermore, those regions that showed significant relations to the apathy scores were also significantly related to the adjective counts in bvFTD in our study. However, adverb production was not related to the degree of apathy in bvFTD. Also, we did not confirm our previous observation that bvFTD patients tend to produce relatively more concrete words than abstract words (Cousins et al., 2017), and this may have been due in part to the limited range of concreteness that could be achieved in a picture displaying many concrete nouns with little evocation of features leading to a description of the picture's abstract characteristics.

It is interesting that adjective counts were negatively correlated with apathy scores in bvFTD, but adverb counts were not. This might be because not all adverbs serve as modifiers in a sentence. For example, so-called pro-adverbs, such as *here* or *there*, perform like function words, replacing prepositional phrases (e.g., *in the kitchen*). It might be the case that bvFTD patients used more pro-adverbs than modifying adverbs, resulting in an insignificant correlation with the apathy score. Additional work is needed to investigate this possibility.

Apathy is not only the most common symptom in bvFTD, occurring in 84% of patients (Rascovsky et al., 2011), but also a prevalent behavioral symptom in patients with other neurodegenerative disorders (e.g., Clark et al., 2008). Our study provided an easily reproducible language variable, adjective production, that might signal the degree of apathy in bvFTD patients. Identifying a language variable that is associated with apathy is particularly valuable, since social/behavioral impairments due to apathy cause the greatest caregiver distress (Massimo et al., 2009). Further study is needed to examine if adjective production is also associated with apathy in patients with other neurodegenerative diseases, such as Alzheimer's disease.

5.4. *Validating an automated lexical analysis of PPA patients' speech*

An important strength of our study is that we were able to validate an automated method for analyzing POS categories in a semi-structured speech sample produced by patients with speech deficits. An automated analysis is reliable in normal, healthy speakers. Here we were able to show that there was over 90% agreement between the automated POS tagging with the Google Universal scheme and the gold-standard POS tagging data of a linguist for speakers with abnormal speech. Indeed, the results of the present study are in line with many previous findings, suggesting that our novel, automated POS tagging and lexical analyses are valid in studying FTD patients' speech.

Speech is central to human daily functioning and our approach has potential to serve as a clinical endpoint for treatment trials. While the present study focuses on cross-sectional data, work in progress assesses objective analyses of our longitudinal speech samples. Language production is a multifaceted process that requires a large expanse of brain tissue and is a sensitive marker for capturing even very early stages of neurodegeneration. Semi-structured speech data such as a picture description is inexpensive to collect on a large scale, when compared to MRI or lumbar puncture for cerebrospinal fluid which are expensive and/or invasive. However, it is nearly impossible to utilize and analyze large-scale speech data in a reproducible manner without an automated method. We believe that the method proposed in this paper can facilitate analyzing large-scale speech data in a quantifiable, automated, and reproducible way and can be used in automatic prescreening for neurodegeneration in the future (e.g., Cho et al., 2020).

6. Conclusion

While our study has many strengths, there are also some limitations that should be kept in mind when interpreting our results. One limitation is that the accuracy of the POS tagging for naPPA patients was not as high as for the other groups. Thus, the results of naPPA patients need to be interpreted with caution. This is an expected result for a POS tagger, since all existing POS taggers are trained with speech/text data of healthy adults. Accuracy could be improved if we trained a POS tagger using our patients' speech samples with speech errors and other abnormalities as a training dataset. Also, since our automated methods rely on texts, there might be, for example, minor speech errors that were transcribed with regular spellings and our pipeline might have missed tagging those tokens as speech errors. We used an open-source POS

tagger in the present study, but we plan to develop NLP tools, including a POS tagger, a syntactic dependency parser, and an automated speech recognition system for automatic transcription that will be trained on patients' speech in the near future. Another limitation is that we had a relatively small number of digitized speech samples and a small number of MRI samples for naPPA patients. This limited our ability to perform statistically robust regression analyses in this patient group. We collect data on a regular basis, and future studies will contain more speech samples.

CRediT author statement

Sunghye Cho: Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization, Naomi Nevler: Conceptualization, Writing – Review & Editing, Visualization, Supervision, Funding acquisition, Sharon Ash: Validation, Investigation, Data Curation, Resources, Writing – Review & Editing, Supervision, Sanjana Shellikeri: Writing – Review & Editing, David J Irwin: Resources, Writing – Review & Editing, Funding acquisition, Lauren Massimo: Resources, Writing – Review & Editing, Funding acquisition, Katya Rascovsky: Resources, Writing – Review & Editing, Christopher Olm: Data Curation, Visualization, Writing – Review & Editing, Mark Liberman: Software, Writing – Review & Editing, Supervision, Funding acquisition, Murray Grossman: Conceptualization, Resources, Writing – Original Draft, Writing – Review & Editing, Supervision, Funding acquisition.

Study funding

National Institutes of Health (AG017586, AG052943, NS088341, DC013063, AG054519, AG066597, AG056054), the Institute on Aging at the University of Pennsylvania, the Alzheimer's Association (AACSF-18-567131), an anonymous donor, and the Wyncote Foundation.

Disclosures

Dr. Grossman participates in clinical trials sponsored by Alector, Eisai and Biogen that are unrelated to this study. He also receives research support from Biogen and Avid that is unrelated to this study, and research support from NIH. Dr. Mark Liberman serves on the Scientific Advisory Board for Baidu Research, USA, and is a co-editor of the Annual Review of Linguistics. All other authors have nothing to disclose.

Appendices

Table A – List of POS categories and mapping between the Google POS tag set and the Penn Treebank tag set. MD, VBD, VBP, and VBZ in the Penn Treebank tags were used to calculate the number of tense-inflected verbs.

Google POS	Penn Treebank	Gloss
NOUN	NN	noun, singular or mass
	NNS	noun, plural
VERB	MD	verb, modal auxiliary
	VB	verb, base form
	VBD	verb, past tense
	VBG	verb, gerund or present participle
	VBN	verb, past participle
	VBP	verb, non-3rd person singular present
	VBZ	verb, 3rd person singular present
ADJ	AFX	affix
(adjective)	JJ	adjective
	JJR	adjective, comparative
	JJS	adjective, superlative
	PRP$	pronoun, possessive
	WDT	wh-determiner (e.g., *which* cookie)
	WP$	wh-pronoun, possessive (e.g., *whose* cookie)
ADV	EX	existential there
(adverb)	RB	adverb
	RBR	adverb, comparative
	RBS	adverb, superlative
	WRB	wh-adverb (e.g., *where*)
PRON	PRP	pronoun
ADP	IN	preposition
X	XX	unknown
INTJ	UH	interjection, exclamation
DET	DT	determiner
CONJ	CC	conjunction

Table B – Demographic and clinical characteristics of the subset of patients with MRI data. The *p*-values for the group differences in this subset were from ANOVA *analyses*, except the sex ratio, where a chi-squared test was used. Linear regression models (all measures but the sex ratio) were used to compare this subset and the remaining dataset to the grand mean of the entire data using the sum coding method. We reported the models' estimated coefficients (β), t-statistics, and *p*-values. A chi-squared test (sex ratio) was used for the comparisons of this subset to the remaining dataset. MMSE: Mini Mental State Exam; BNT: Boston Naming Test; PPT: Pyramids and Palm Trees Test; F: females; M: males.

	Controls (n = 18)	bvFTD (n = 42)	naPPA (n = 8)	svPPA (n = 26)	Group differences in this subset	Comparison with the full set
Age (years)	65.9 (6.8)	63 (8.5)	65.5 (8.1)	61.2 (7.1)	$F(3,90) = 1.53$, $p = .21$	$\beta = 2.5$, $t(167) = 1.91$, $p = .059$
Sex	9 F, 9 M	15 F, 27 M	2 F, 6 M	17 F, 9 M	$\chi = 7.26$, $p = .06$	$\chi = .24$, $p = .62$
Education (years)	16.1 (2.9)	15.9 (2.2)	17.4 (3)	15.3 (2.6)	$F(3,90) = 1.37$, $p = .26$	$\beta = -.16$, $t(167) = -.35$, $p = .72$
Disease duration (years)	–	4 (3.4)	3 (2)	3.6 (2)	$F(2,73) = .43$, $p = .66$	$\beta = .48$, $t(132) = 1.4$ $p = .16$
MMSE (0–30)	28.9 (1.1)	25.1 (4.3)	25.1 (3.4)	23.6 (6.1)	$F(3,88) = 5.26$, $p = .002$	$\beta = .26$, $t(149) = .28$ $p = .78$
BNT (0–30)	27.7 (2.7)	24.5 (4.1)	24.8 (5.1)	7.7 (6.3)	$F(3,89) = 90.81$, $p < .001$	$\beta = .44$, $t(139) = .32$ $p = .75$
PPT (0–52)	51.3 (1.1)	45.4 (6.9)	48.5 (3.7)	39.1 (7.1)	$F(3,48) = 8.75$, $p < .001$	$\beta = -.52$, $t(71) = -.37$, $p = .71$
Animals and tools (max 60 sec)	16.8 (5)	10 (4.9)	9.8 (4.8)	6 (3.9)	$F(3,86) = 18.52$, $p < .001$	$\beta = .18$, $t(135) = .15$, $p = .88$

Table C – POS *counts* per 100 words and lexical measures of the subset of patients with MRI data.

	Controls	bvFTD	naPPA	svPPA
Nouns	19.42 (4.67)	21.67 (6.94)	23.65 (7.33)	17.43 (5.12)
Unique nouns	14.4 (3.37)	16.26 (6.27)	19.44 (3.84)	12.24 (4.6)
Pronouns	7.64 (2.33)	6.21 (3.58)	5.55 (1.86)	9.4 (3.89)
wh-words	.63 (.35)	1.3 (3.04)	1.16 (1.81)	.9 (1.32)
Tense-inflected verbs	12.02 (1.56)	12.26 (3.8)	11.28 (3.88)	13.71 (3.2)
Verbs	22.46 (3.1)	22.67 (4.56)	20.81 (4.67)	24.11 (4.68)
Speech errors/Partial words	.81 (1.16)	1.17 (1.86)	3.36 (3.93)	.73 (1.12)
Adverbs	5.61 (1.79)	5.46 (3.31)	3.51 (2.88)	7.94 (4.69)
Adjectives	6.01 (1.68)	3.89 (2.38)	3.35 (2.28)	3.3 (2.6)
Prepositions	10.81 (1.52)	8.34 (4.07)	5.48 (2.73)	7.78 (3.96)
Total words	194.22 (75.56)	112.23 (67.5)	85.75 (50)	121.88 (66.49)
Determiners	13.6 (2.16)	15.6 (3.93)	16.5 (4.24)	12.76 (5.3)
Conjunctions	4.38 (1.82)	5.01 (2.78)	4.36 (3.24)	5.02 (3.31)
Interjections	5.02 (2.43)	5.7 (3.83)	8.9 (4.78)	6.45 (5.43)
Ratio of content to function words	1.31 (.26)	1.36 (.35)	1.28 (.24)	1.32 (.32)
Abstractness (noun)	1.54 (.24)	1.48 (.26)	1.35 (.21)	1.86 (.51)
Ambiguity (noun)	1.69 (.05)	1.66 (.06)	1.63 (.09)	1.77 (.13)
Frequency (noun)	3.58 (.17)	3.61 (.28)	3.49 (.4)	4.01 (.44)
Familiarity (noun)	2.36 (.03)	2.35 (.05)	2.36 (.03)	2.41 (.07)
AoA (noun)	4.4 (.38)	4.21 (.42)	4.14 (.5)	4.1 (.46)
Cross entropy	9.75 (.52)	9.66 (.74)	10.21 (1)	9.18 (.58)
Lexical diversity	.85 (.04)	.79 (.09)	.8 (.06)	.8 (.1)

REFERENCES

Adlam, A. L. R., Bozeat, S., Arnold, R., Watson, P., & Hodges, J. R. (2006). Semantic knowledge in mild cognitive impairment and mild Alzheimer's disease. *Cortex, 42*(5), 675–684. https://doi.org/10.1016/S0010-9452(08)70404-0

Amici, S., Ogar, J., Brambati, S. M., Miller, B. L., Neuhaus, J., Dronkers, N. L., et al. (2007). Performance in specific language tasks correlates with regional volume changes in progressive aphasia. *Cognitive and Behavioral Neurology, 20*(4), 203–211. https://doi.org/10.1097/WNN.0b013e31815e6265

Ash, S., Evans, E., O'Shea, J., Powers, J., Boller, A., Weinberg, D., et al. (2012). Differentiating primary progressive aphasias in connected speech production. *Neurology, 34*, 246.

Ash, S., McMillan, C., Gross, R. G., Cook, P., Gunawardena, D., Morgan, B., et al. (2013). Impairments of speech fluency in Lewy body spectrum disorder. *Brain and Language, 120*(3), 290–302. https://doi.org/10.1016/j.bandl.2011.09.004

Ash, S., McMillan, C., Gunawardena, D., Avants, B., Morgan, B., Khan, A., et al. (2010). Speech errors in progressive non-fluent aphasia. *Brain and Language, 113*, 13–20.

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., & Grossman, M. (2006). Trying to tell a tale. *Neurology, 66*(9), 1405–1413.

Ash, S., Moore, P., Vesely, L., Gunawardena, D., McMillan, C., Anderson, C., et al. (2009). Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics, 22*(4), 370–383. https://doi.org/10.1016/j.jneuroling.2008.12.001

Ash, S., Ternes, K., Bisbing, T., Min, N. E., Moran, E., York, C., et al. (2016). Dissociation of quantifiers and object nouns in speech in focal neurodegenerative disease. *Neuropsychologia, 89*(1), 141–152. https://doi.org/10.1016/j.neuropsychologia.2016.06.013

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

Bonner, M., Price, A. R., Peelle, J. E., & Grossman, M. (2016). Semantics of the visual environment encoded in

parahippocampal cortex. *Journal of Cognitive Neuroscience, 28*(3), 361–378.
Bonner, M., Vesely, L., Price, C., Anderson, C., Richmond, L., Farag, C., et al. (2009). Reversal of the concreteness effect in semantic dementia. *Cognitive Neuropsychology, 26*(6), 568–579. https://doi.org/10.1080/02643290903512305
Breedin, S., Saffran, E., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology, 11*(6), 617–660.
Brysbaert, M., Mandera, P., & Keuleers, E. (2018). Word prevalence norms for 62 ,000 English lemmas. *Behavior Research Methods, 51*(2), 467–479.
Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977
Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5
Cho, S., Nevler, N., Shellikeri, S., Ash, S., Liberman, M., & Grossman, M. (2020). Automatic classification of primary progressive aphasia patients using lexical and acoustic features. In *Proceedings of language Resources and evaluation conference (LREC) 2020 workshop on Resources and processing linguistic, para-linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID-3)* (pp. 60–65).
Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004). *Fisher English training speech corpus*. Philadelphia: Linguistic Data Consortium.
Clark, D. E., van Reekum, R., Simard, M., Streiner, D. L., Conn, D., Cohen, T., et al. (2008). Apathy in dementia: Clinical and sociodemographic correlates. *The Journal of Neuropsychiatry and Clinical Neurosciences, 20*, 337–347.
Conover, W. (1980). *Practical nonparametric statistics* (2nd ed.). New York: John Wiley.
Cousins, K., Ash, S., Irwin, D. J., & Grossman, M. (2017). Dissociable substrates underlie the production of abstract and concrete nouns. *Brain and Language, 165*, 45–54. https://doi.org/10.1016/j.bandl.2016.11.003
Cousins, K., Ash, S., Olm, C. A., & Grossman, M. (2018). Longitudinal changes in semantic concreteness in semantic variant primary progressive aphasia (svPPA). *eNeuro, 5*(6), 1–10. https://doi.org/10.1523/ENEURO.0197-18.2018
Cousins, K. A., York, C., Bauer, L., & Grossman, M. (2016). Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia, 84*, 244–251. https://doi.org/10.1016/j.neuropsychologia.2016.02.025
Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*, 94–100.
Croot, K., Ballard, K., Leyton, C. E., & Hodges, J. R. (2012). Apraxia of speech and phonological errors in the diagnosis of nonfluent/agrammatic and logopenic variants of primary progressive aphasia. *Journal of Speech, Language, and Hearing Research, 55*(5), 1562–1572. https://doi.org/10.1044/1092-4388(2012/11-0323
Cunningham, K. T., & Haley, K. L. (2020). Analysis in Aphasia: Moving-average type-token ratio and word information measure. *Journal of Speech, Language, and Hearning Research, 63*(3), 710–721.
Daducci, A., Gerhard, S., Griffa, A., Lemkaddem, A., Cammoun, L., Gigandet, X., et al. (2012). The connectome mapper: An open-source processing pipeline to map connectomes with MRI. *Plos One, 7*(12). https://doi.org/10.1371/journal.pone.0048121
Das, S., Avants, B. B., Grossman, M., & Gee, J. C. (2009). Registration based cortical thickness measurement. *Neuroimage, 45*(3), 867–879. https://doi.org/10.1016/j.neuroimage.2008.12.016
Farag, C., Troiani, V., Bonner, M., Powers, C., Avants, B., Gee, J., et al. (2010). Hierarchical organization of scripts: Converging evidence from fmri and frontotemporal degeneration. *Cerebral Cortex, 20*(10), 2453–2463. https://doi.org/10.1093/cercor/bhp313
Godfrey, J., & Holliman, E. (1997). *Switchboard-1 release 2*. Philadelphia: Linguistic Data Consortium.
Goodglass, H., Kaplan, E., & Weintraub, S. (1983). *Boston diagnostic aphasia examination*. Philadelphia, PA: Lea & Febiger.
Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., et al. (2004). Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology, 55*, 335–346.
Gorno-Tempini, M. L., Hillis, A., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology, 76*(11), 1006–1014.
Grossman, M. (2012). The non-fluent/agrammatic variant of primary progressive aphasia. *The Lancet Neurology, 11*(6), 545–555. https://doi.org/10.1016/S1474-4422(12)70099-6
Grossman, M., Mickanin, J., Onishi, K., Hughes, E., D'Esposito, M., Ding, X. S., et al. (1996). Progressive nonfluent aphasia: Language, cognitive, and PET measures contrasted with probable Alzheimer's disease. *Journal of Cognitive Neuroscience, 8*(2), 135–154. https://doi.org/10.1162/jocn.1996.8.2.135
Grossman, M., Rhee, J., & Moore, P. (2005). Sentence processing in frontotemporal dementia. *Cortex, 41*(6), 764–777. https://doi.org/10.1016/S0010-9452(08)70295-8
Gunawardena, D., Ash, S., McMillan, C. T., Avants, B., Gee, J., & Grossman, M. (2010). Why are patients with progressive nonfluent aphasia nonfluent? *Neurology, 75*(7), 588–594.
Hardy, C., Buckley, A. H., Downey, L. E., Lehmann, M., Zimmerer, V. C., Varley, R. A., et al. (2016). The language profile of behavioral variant frontotemporal dementia. *Journal of Alzheimer's Disease, 50*(2), 359–371. https://doi.org/10.3233/JAD-150806
Hillis, A., Oh, S., & Ken, L. (2004). Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology, 55*, 268–275.
Hillis, A., Tuffiash, E., & Caramazza, A. (2002). Modality-specific deterioration in naming verbs in nonfluent primary progressive aphasia. *Journal of Cognitive Neuroscience, 14*(7), 1099–1108. https://doi.org/10.1162/089892902320474544
Hodges, J., & Patterson, K. (2007). Semantic dementia: A unique clinicopathological syndrome. *Lancet Neurology, 6*, 1004–1014.
Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x
Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *EMNLP 2015: Conference on empirical methods in natural language processing* (pp. 1373–1378). https://doi.org/10.18653/v1/d15-1162
Howard, D., & Patterson, K. (1992). *Pyramids and Palm Trees: A test of semantic access from pictures and words*. Bury St. Edmunds, UK: Thames Valley Test Company.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84.
Josephs, K., Duffy, J. R., Strand, E. A., Machulda, M. M., Senjem, M. L., Lowe, V. J., et al. (2013). Syndromes dominated by apraxia of speech show distinct characteristics from agrammatic PPA. *Neurology, 81*, 337–345.

Josephs, K., Duffy, J. R., Strand, E. A., MacHulda, M. M., Senjem, M. L., Master, A. V., et al. (2012). Characterizing a neurodegenerative syndrome: Primary progressive apraxia of speech. *Brain, 135*(5), 1522–1536. https://doi.org/10.1093/brain/aws032

Josephs, K., Duffy, J. R., Strand, E. A., Whitwell, J. L., Layton, K. F., Parisi, J. E., et al. (2006). Clinicopathological and imaging correlates of progressive aphasia and apraxia of speech. *Brain, 129*(6), 1385–1398. https://doi.org/10.1093/brain/awl078

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test. Austin, TX: Pro-Ed*.

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

Kramer, J., Jurik, J., Sha, S. J., Rankin, K. P., Rosen, H. J., Johnson, J. K., et al. (2003). Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. *Cognitive and Behavioral Neurology, 16*(4), 211–218. https://doi.org/10.1097/00146965-200312000-00002

Lenth, R. V. (2016). Least-square means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1–33.

Lezak, M., Howieson, D. B., & Loring, D. W. (1983). *Neuropsychological assessment*. New York: Oxford University Press.

Libon, D. J., Rascovsky, K., Gross, R. G., White, M. T., Xie, S. X., Dreyfuss, M., et al. (2011). The Philadelphia Brief assessment of cognition (PBAC): A validated screening measure for dementia. *Clinical Neuropsychologist, 25*(8), 1314–1330.

Libon, D. J., Xie, S. X., Moore, P., Farmer, J., Antani, S., McCawley, G., et al. (2007). Patterns of neuropsychological impairment in frontotemporal dementia. *Neurology, 68*(5), 369–375. https://doi.org/10.1212/01.wnl.0000252820.81313.9b

Macoir, J. (2009). Is a plum a memory problem?. Longitudinal study of the reversal of concreteness effect in a patient with semantic dementia. *Neuropsychologia, 47*(2), 518–535. https://doi.org/10.1016/j.neuropsychologia.2008.10.006

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

Marcus, D., Wang, T., Parker, J., Csernansky, J., Morris, J., & Buckner, R. (2007). Cross-sectional MRI data in young, middle ages, nondemented, and demented older adults. *Journal of Cognitive Neuroscience, 19*(9), 1498–1507.

Massimo, L., Powers, J. P., Evans, L. K., McMillan, C. T., Rascovsky, K., Eslinger, P., et al. (2015). Apathy in frontotemporal degeneration: Neuroanatomical evidence of impaired goal-directed behavior. *Frontiers in Human Neuroscience, 9*, 611.

Massimo, L., Powers, C., Moore, P., Vesely, L., Avants, B., Gee, J., et al. (2009). Neuroanatomy of apathy and disinhibition in frontotemporal lobe degeneration. *Dementia and Geriatric Cognitive Disorders, 27*, 96–104.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–337.

Moscoso del Prado Martín. (2017). Vocabulary, grammar, sex, and aging. *Cognitive Science, 41*, 950–975.

Nevler, N., Ash, S., Irwin, D. J., Liberman, M., & Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology, 6*(1), 4–14. https://doi.org/10.1002/acn3.653

Ogar, J., Dronkers, N. F., Brambati, S. M., Miller, B. L., & Gorno-Tempini, M. L. (2007). Progressive nonfluent aphasia and its characteristic motor speech deficits. *Alzheimer Disease and Associated Disorders, 21*(4), S23–S30. https://doi.org/10.1097/WAD.0b013e31815d19fe

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–987. https://doi.org/10.1038/nrn2277

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the international conference on language Resources and evaluation* (pp. 2089–2096).

R Core Team. (2019). *R: a language and environment for statistical computing*.

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain, 134*(9), 2456–2477. https://doi.org/10.1093/brain/awr179

Rhee, J., Antiquena, P., & Grossman, M. (2001). Verb comprehension in frontotemporal degeneration: The role of grammatical, semantic and executive components. *Neurocase, 7*(2), 173–184. https://doi.org/10.1093/neucas/7.2.173

Rogers, T. T., Patterson, K., Jefferies, E., & Lambon Ralph, M. A. (2015). Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. *Neuropsychologia, 76*, 220–239. https://doi.org/10.1016/j.neuropsychologia.2015.04.015

RStudio Team. (2016). *RStudio*. Boston, MA: Integrated Development for R.

Slegers, A., Filiou, R.-P., Montembeault, M., & Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease: A systematic review. *Journal of Alzheimer's disease, 65*(2), 519–524.

Strand, E., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders, 51*, 43–50. https://doi.org/10.1016/j.jcomdis.2014.06.008

Tappen, R. M., Williams, C. L., Barry, C., & DiSesa, D. (2002). Conversation intervention with Alzheimer's Patients: Increasing the relevance of communication. *Clinical Gerontology, 24*(3–4), 63–75.

Thompson, C. K., & Mack, J. E. (2014). Grammatical impairments in PPA. *Aphasiology, 28*(8–9), 1018–1037.

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., et al. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage, 99*, 166–179. https://doi.org/10.1016/j.neuroimage.2014.05.044

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323–352.

Weintraub, S., Rubin, N. P., & Mesulam, M.-M. (1990). Primary progressive aphasia: Longitudinal course, neuropsychological profile, and language features. *Archives of Neurology, 47*(12), 1329–1335. https://doi.org/10.1001/archneur.1990.00530120075013

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., et al. (2013). *OntoNotes release 5.0*. Philadelphia: Linguistic Data Consortium.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., et al. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain, 113*(7), 2069–2088. https://doi.org/10.1093/brain/awq129

Automatic Classification of Primary Progressive Aphasia Patients Using Lexical and Acoustic Features

Sunghye Cho[1], Naomi Nevler[2], Sanjana Shellikeri[2], Sharon Ash[2], Mark Liberman[1], Murray Grossman[2]

Linguistic Data Consortium[1], Penn Frontotemporal Degeneration Center[2]
University of Pennsylvania, 3400 Walnut Street, Philadelphia, PA, USA

{csunghye, myl}@ldc.upenn.edu
{naomine, Sanjana.Shellikeri, ash, mgrossma}@pennmedicine.upenn.edu

Abstract

Two variants of primary progressive aphasia (PPA) are subtypes of frontotemporal degeneration (FTD), which is the most common type of dementia among individuals under 60 years of age. Semantic variant PPA (svPPA) patients present with semantic deficits in single word use, whereas nonfluent/agrammatic PPA (naPPA) patients produce simplified speech with frequent speech errors and slow speech rates. In this study, we built machine learning systems to classify PPA patients (n=63) and healthy elderly controls (n=36). We automatically extracted 18 lexical and 21 acoustic features with a natural language processing library and a speech activity detector, and we trained classifiers, experimenting with various feature selection and reduction techniques. Our models showed high accuracy, correctly distinguishing patients from controls in more than 90% of cases, svPPA patients from naPPAs in about 89% of cases, and controls, svPPA, and naPPA patients in more than 80% of cases. Our results show that classification of PPA patients using automatically derived linguistic features from digitized speech samples is highly promising, and could potentially be applied in community settings for prescreening. We plan to extend this project by including more features and additional FTD subgroups in the near future.

Keywords: Primary progressive aphasia, automatic classification, narrative speech

1. Introduction

Frontotemporal degeneration (FTD) is a type of focal dementia caused by atrophy in the brains' frontal and temporal lobes. It is the most common type of neurodegenerative disease among people under 60 years of age (Ratnavalli et al., 2002). Since individuals diagnosed with FTD are relatively young, usually still in the workforce, the personal and societal costs of the disease are substantial. For example, FTD diagnosis often results in early departure from the workforce, increasing economic burden for a household with an FTD patient and negatively affecting not only patients but also the quality of life of their families (Galvin et al., 2017). Because there are no disease-modifying drugs approved for FTD, earlier screening and slowing the apparent disease progression rate through behavioral adjustments to the environment are key to helping patients and their families. This paper proposes three machine learning systems to automatically classify two subgroups of FTD that could potentially be applied in prescreening.

About half of patients with FTD present with a linguistic impairment known as primary progressive aphasia (PPA), and sometimes this can be accompanied by a social-behavioral impairment known as behavioral variant FTD (bvFTD). There are several variants of PPA. Among these subgroups, semantic variant PPA (svPPA) patients are characterized by impaired confrontation naming, frequent substitution of pronouns for nouns, and difficulty in processing concrete words, although they show intact prosody and syntax (e.g., Amici et al., 2007; Bonner et al., 2016; Cousins et al., 2016; Nevler et al., 2019). Nonfluent/agrammatic PPA (naPPA) patients, on the other hand, present with effortful speech, slow speech rates, frequent speech errors, simplified grammar, and difficulty in retrieving verbs (e.g., Ash et al., 2009; Grossman et al., 1996; Rhee et al., 2001). Patients with either of the two subtypes have frontotemporal lobar degeneration spectrum pathology, which is commonly associated with misfolding of TDP-43 or tau proteins.

Since PPA patients show salient linguistic characteristics, we would expect automatic classification by means of linguistic features to yield high levels of accuracy. There are a few previous studies that have pursued this approach. Fraser et al. (2014) extracted 58 lexical and semantic features from the speech samples of 10 svPPA and 14 naPPA patients and 16 controls. The authors trained classifiers only with significant features for three different tasks: control versus svPPA, control versus naPPA, and svPPA versus naPPA. Their models for controls versus svPPA/naPPA showed high levels of accuracy, from 90% to 100%. However, their best performance for classifying svPPA and naPPA patients was only 79.2% accurate, suggesting that classifying patient groups is more difficult than distinguishing patients from controls. Peintner et al. (2008) extracted 41 acoustic, 81 LIWC (Language Inquiry and Word Count; Pennebaker et al., 2001), and 11 lexical features from 39 participants (9 bvFTD, 8 naPPA, 13 svPPA, and 9 controls), and trained classifiers for various classification tasks. Their composite feature set (significant features from each feature set) showed accuracy over 90% in most classification tasks, except control versus bvFTD and four-way classification. However, they did not list what features were used in the composite set, making it difficult to reproduce their results. Themistocleous et al. (2019) extracted 14 acoustic features, such as mean fundamental frequency and amplitude differences between the first and second harmonics, from 50 patients (17 logopenic variant PPA (lvPPA), 14 svPPA, 11 naPPA, and 8 naPPA with apraxia of speech) and trained classifiers with 3-fold group cross validations and a one-against-all strategy. Their models correctly identified naPPA 82% of the time and svPPA 66% of the time. The authors only used acoustic features, which explains why the accuracy of svPPA patients, who rarely show impairments in prosody, was relatively low. More importantly, all previous studies have had relatively small datasets, raising

the question of whether their results could be generalized to larger datasets. In this paper, we studied 99 participants (63 patients and 36 controls) to investigate whether lexical and acoustic features could predict the diagnostic status of the participants.

2. Objectives

Our objectives were to train three predictive models for classifying (1) controls vs. patients, (2) svPPA vs. naPPA patients, and (3) controls, svPPA and naPPA patients, experimenting with different feature selection and reduction techniques, and to identify predictive features for classifying PPA patients.

3. Methods

3.1 Participants

Our participants consisted of 63 patients diagnosed clinically with either svPPA or naPPA and 36 healthy elderly controls. Forty-two of the 63 patients had svPPA and 21 were naPPA patients. The patients were diagnosed by experienced neurologists at the Department of Neurology of the Hospital of the University of Pennsylvania in accordance with published criteria (Gorno-Tempini et al., 2011). Of the 42 svPPA patients, 32 showed concomitant mild behavioral symptoms, which is a common co-occurrence in this group. We focused on frontotemporal lobar degeneration (FTLD) spectrum pathology in this study, and so we did not include lvPPA patients, who most often have Alzheimer's pathology. Our participants were matched on sex ratio and education levels, but not on age, because naPPA patients on average have an later disease onset than svPPA patients (Johnson et al., 2005). The patient groups did not differ on the Mini Mental State Exam scores (MMSE) or disease durations, but they significantly differed on the Boston Naming Test (BNT) scores, which is expected due to svPPA patients' difficulty in naming tasks. All participants were native speakers of English. The study was approved by the Institutional Review Board of the Hospital of the University of Pennsylvania, and all participants signed a written consent form. Participants' demographic and neuropsychological characteristics are summarized in Table 1.

	controls	**svPPA**	**naPPA**	***p*-value**
Age	68.5 (7.9)	63.3 (6.9)	70.4 (9.4)	0.001
Sex	23 F/13 M	23 F/19 M	11 F/11 M	0.483
Education (years)	15.9 (2.5)	15.1 (2.8)	15.3 (3.1)	0.408
MMSE (range: 0-30)	29.2 (1)	22.1 (6.3)	22.7 (5.9)	< 0.001
BNT (range: 0-30)	27.9 (2.5)	7.5 (6.4)	24.7 (4.6)	< 0.001
Disease duration (yrs)	NA	3.9 (2)	3.2 (1.9)	0.214
Total number of words in Cookie Theft	174.4 (66.4)	148.1 (62.8)	91.0 (55.8)	< 0.001

Table 1: Mean (SD) demographic and neuropsychological characteristics of the participants. MMSE: Mini Mental State Exam, BNT: Boston Naming Test.

3.2 Data

The Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 1983) was used to elicit narrative speech from the participants. Participants described the picture for about one minute, and their descriptions were digitally recorded. Some patients made several recordings, but we used the earliest recording of each participant in this analysis in order to differentiate among the conditions early in the disease course. An expert linguist generated verbatim transcription of the picture descriptions, including all non-verbal speech, hesitations and false starts, and a team of trained annotators at the Linguistic Data Consortium (LDC) of the University of Pennsylvania reviewed and revised the annotations for quality checking.

4. Feature Extraction

4.1 Lexical Features

We ran a POS tagger in spaCy (Honnibal & Johnson, 2015) to automatically tag POS categories of all words that the participants produced in the picture descriptions. Before running the tagger, we cleaned the transcripts by removing interviewers' prompts and annotations for non-verbal speech. A professional linguist manually validated the accuracy of spaCy with a subset of our data (n=21). The mean group accuracy varied from 95% (controls) to 90% (PPAs). There was no significant difference in the accuracy among patient groups (p>0.05). Since the accuracy of the spaCy POS tagger with their basic model ('en_core_web_sm') was high, we did not train a POS tagger separately in this study. The POS tokens were tallied per participant, and the count of each POS category per 100 words was calculated (= (raw counts/total number of words) * 100). In addition to the frequency of each POS category, we measured the number of tense-inflected verbs and unique nouns per 100 words. We summed the number of modal auxiliary verbs, past tense verbs and present tense verbs that spaCy tagged to count the number of tense-inflected verbs per 100 words. The number of noun lemmas was used for the number of unique nouns per 100 words.

We also rated nouns that participants produced for concreteness (Brysbaert et al., 2014), semantic ambiguity (Hoffman et al., 2013), word frequency (Brysbaert & New, 2009), age of acquisition (AoA; Brysbaert et al., 2018) and word familiarity (Brysbaert et al., 2018) for their potential to distinguish svPPA patients from others. Since the published norms we used had a limited number of words, we rated the lemma of a noun if a noun itself was not listed in the published norms. A noun was not rated if neither the noun nor its lemma was listed in the norms. In total, we had 18 text-related features: POS counts per 100 words (nouns, verbs, adjectives, adverbs, prepositions, determiners, conjunctions, interjections, pronouns, and speech errors/partial words—[X] in spaCy), number of tense-inflected verbs and unique nouns per 100 words, lexical features of nouns (concreteness, ambiguity, frequency, AoA, familiarity), and total number of words.

4.2 Acoustic Features

We used an in-house Gaussian Mixture Models-Hidden Markov Models based Speech Activity Detector (SAD) developed at the LDC to segment the recordings into speech and silent pause segments. We set the minimum duration of a speech segment at 250 ms and that of a silence segment at 150 ms. We reviewed the outputs of SAD, corrected wrong segmentations, and excluded interviewers' speech and non-verbal speech segments. Using the durations of speech and silent pause segments, we extracted 12 durational features:

- The mean duration of speech and pause segments
- The number of total pauses and speech segments
- Total speech time (speech only)
- Total pause time (pause only)
- Total time (speech time + pause time)
- Sample duration (duration of the entire recording)
- Percent of speech time of the total time
- Breath frequency (= number of pauses over total time)
- Speech frequency (= number of speech segments over total time)
- Pause rate per minute (= number of pauses over total speech time)

We also pitch-tracked speech segments of the participants with a script in Praat (Boersma & Weenink, 2020) and extracted the 10^{th} to 90^{th} fundamental frequency (f0) percentile values for each speaker. To minimize individual differences in pitch due to physiological factors, such as sex, height, and the size of the larynx, the extracted f0 values in Hz were converted to semitones (St) using each speaker's 10^{th} percentile as a baseline: $St = \log_2(Hz / 10^{th} \text{ percentile})*12$. We had 21 acoustic features in total, including pitch percentile values along with the 12 durational features. The final feature set included 18 lexical and 21 acoustic features and 3 demographic characteristics of the participants: age, sex, and education level.

5. Model Training

We trained two different machine learning algorithms from the scikit-learn package (Pedregosa et al., 2011) in Python: Random Forest and Support Vector Machine (SVM) classifiers. In all models, we imputed missing values with a mean value using SimpleImputer and standardized features with StandardScaler in scikit-learn for effective learning. We performed leave-one-out cross-validation (CV) to evaluate the generalizability of the models and reported the average accuracy of all CV folds.

We experimented with feature selection and reduction methods. For feature selection, we performed t-tests (for binary classifications) and trained models with features that were significant at the level of $p < 0.05$, 0.01, 0.005, and 0.001. We used the same feature set used in the control-patient pairwise classification for the three-way classification (control vs. svPPA vs. naPPA). We compared the performance of models trained with selected features and a model without any feature selection. For feature reduction, we performed Principal Component Analysis (PCA) and trained models, varying the number of components from 1 to 10. We compared the performance of models trained with PCA components and that of a model trained without any feature reduction and reported the best performance after tuning hyperparameters.

6. Classification Results

6.1 Binary Classification between Controls and Patients

An SVM classifier trained with all features which were reduced to 10 PCA components performed best in this classification task, showing 90.9% accuracy and 0.94 area under the curve (AUC). Our model correctly identified 33 controls out of 36 and 57 patients out of 63. The full classification report is shown in Table 2, and the receiver operating characteristic (ROC) curve for this contrast is provided in Figure 1.

	Accuracy	Precision	Recall	F1-score
Controls	0.92	0.85	0.92	0.88
Patients	0.90	0.95	0.90	0.93
Weighted average	0.91	0.91	0.91	0.91

Table 2: Classification report of the SVM classifier for the classification of patients and controls.

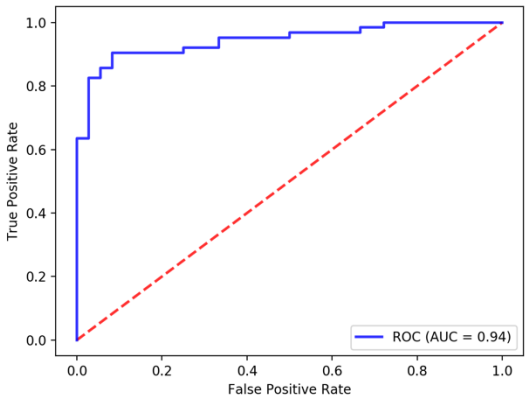


Figure 1: Receiver Operator Characteristic Curve for the classification of controls and patients.

6.2 Binary Classification of Patient groups

A Random Forest classifier trained with features that were significant at the level of $p<0.005$ and reduced to three PCA components performed best in this classification task. The model showed 88.9% accuracy with 0.87 AUC. The model correctly identified 40 svPPA patients out of 42 and 16 naPPA patients out of 21. Our model resulted in a higher F1-score for classifying svPPA patients (0.92) than naPPA patients (0.82), suggesting that in general identifying naPPA patients was more difficult than identifying svPPA patients. The full classification scores are in Table 3, and the ROC curve for this contrast is provided in Figure 2.

	Accuracy	Precision	Recall	F1-score
svPPA	0.95	0.89	0.95	0.92
naPPA	0.76	0.89	0.76	0.82
Weighted average	0.89	0.89	0.89	0.89

Table 3: Classification report of the Random Forest classifier for the classification of svPPA and naPPA patients.

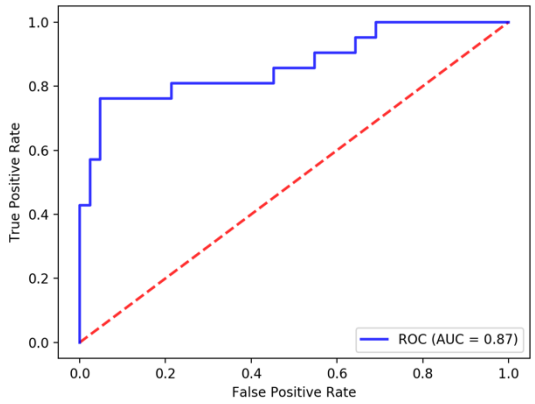


Figure 2: Receiver Operator Characteristic Curve for the classification of svPPA and naPPA patients.

The features that were selected included counts of nouns, pronouns, verbs, tense-inflected verbs, speech errors/partial words, unique nouns per 100 words; concreteness, semantic ambiguity, frequency of nouns; participants' age and total number of pauses. Figure 3 shows group differences in the selected features.

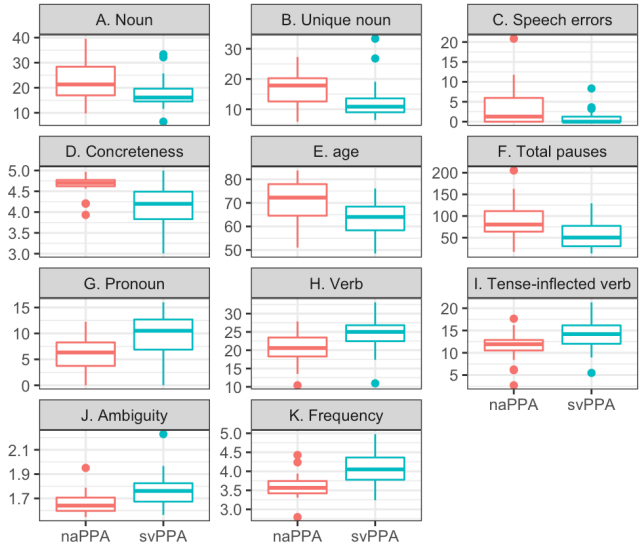


Figure 3: Group differences in selected features for the classification of svPPA and naPPA patients. The POS counts and the numbers of tense-inflected verbs and unique nouns are per 100 words. The top two rows show features where values of naPPA patients are significantly higher than those of svPPA and the bottom two rows show features where values of svPPA patients are significantly higher than those of naPPA (both at $p<0.005$).

Among the 11 selected features, most were lexical, and only one acoustic feature, total number of pauses, was selected. As expected, semantic aspects of nouns that patients produced, such as concreteness and semantic ambiguity, were important features in distinguishing svPPA patients from naPPA patients. Further discussion of the acoustic features in PPA patients can be found in Nevler et al. (2019), and further discussion of the lexical features can be found in Cho et al. (under review).

6.3 Three-way Classification

An SVM classifier trained with all features without any feature reduction performed best for the three-way classification, yielding 80.8% accuracy with 0.9 macro-averaged AUC. The model correctly identified 32 controls out of 36, 34 svPPA patients out of 42, and 14 naPPA patients out of 21. The model's F1-score is high for controls and svPPA patients (> 0.8), but it was below 0.7 for naPPA patients, again suggesting that naPPA patients were difficult to identify. The full classification report and the confusion matrix are provided in Tables 4 and 5, and the ROC curve for this contrast is provided in Figure 4.

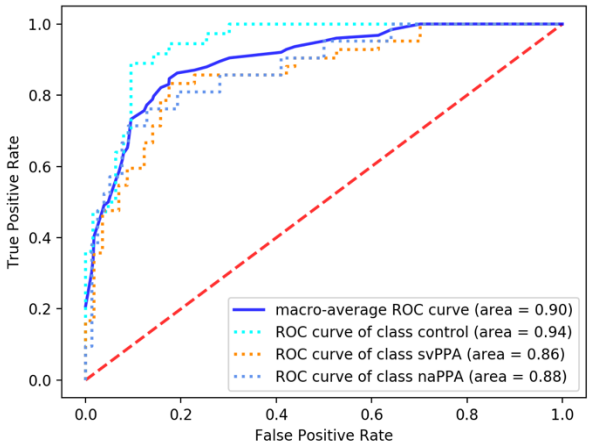


Figure 4: Receiver Operator Characteristic Curve for the classification of controls and svPPA and naPPA patients.

	Accuracy	Precision	Recall	F1-score
Control	0.89	0.84	0.89	0.86
svPPA	0.81	0.83	0.81	0.82
naPPA	0.67	0.70	0.67	0.68
Weighted average	0.81	0.81	0.81	0.81

Table 4: Classification report of the SVM classifier for the three-way classification.

	Control	svPPA	naPPA
Controls	32	2	2
svPPA	4	34	4
naPPA	2	5	14

Table 5: Confusion matrix of the three-way classification (column: actual, row: predicted). The number of accurately classified participants is highlighted in gray.

7. Discussion and Conclusion

This paper reports the results of automatic classification systems for three classification tasks: i) control versus patients, ii) svPPA versus naPPA patients, and iii) control versus svPPA versus naPPA. We automatically extracted 18 lexical features from one-minute narrative speech samples using spaCy, one of the most modern, state-of-the-art natural language processing libraries in Python. We also automatically extracted 21 acoustic and durational features with SAD. Using these features with additional demographic information, we trained three machine learning classifiers, experimenting with different feature selection and reduction techniques, and used leave-one-out cross-validation. We found group differences in the selected features. Our model for the control versus patient classification trained with all features, which were reduced to 10 PCA components, correctly distinguished patients from controls in more than 90% of cases. Our classifier for the svPPA versus naPPA task selected 11 features (9 lexical, 1 acoustic and 1 demographic), which were later reduced to 3 PCA components. Our classifier correctly identified the diagnostic group of the patients with 88.9% accuracy, which outperformed the system for the same task in previous studies (79.2% in Fraser et al., 2014; 82% for naPPA patients in Themistocleous et al., 2019). Lastly, our system for the three-way classification, which was trained with all features without any feature reduction, showed high overall accuracy (over 80%) in classifying controls, svPPA and naPPA patients, which is much higher than the chance level (33.3%). The performance of the systems in this report is highly promising in that we only had one-minute narrative speech samples, which are quick and easy to collect. We believe that this line of research could potentially benefit populations with the earliest features of PPA.

Our models performed well, but there is still room for improvement, in particular, for the three-way classification system, where classification of naPPA was < 80%. In the future, we plan to include more features, such as letter or category fluency scores, Mel-frequency cepstral coefficients, or word frequency as log-odds ratio (Monroe et al., 2008) to improve the performance of the models. We also aim to extend our research by including more patient groups. First, we would consider evaluating patients with lvPPA, which is another variant of PPA associated with Alzheimer's disease pathology, with frequent filler words (*um* or *uh*) as a prominent feature. Second, we would consider including bvFTD patients, who have pathology similar to that of svPPA and naPPA patients. Although without obvious aphasia, these patients do have subtle speech deficits (Nevler et al., 2018). In addition, we plan to collect conversational data in the near future to explore subtle group differences among these patient groups that may not have been captured in monologue, narrative speech samples. In natural conversation, speakers employ a variety of prosodic features to deliver the intended message effectively. We believe these additional features will improve the models' performance.

8. Acknowledgements

We thank the participants and their family members for participating in the study and the research assistants who helped collect the data. This study was funded by National Institutes of Health (AG017586, AG053940, AG052943, NS088341, DC013063, AG054519), the Institute on Aging at the University of Pennsylvania, the Alzheimer's Association (AACSF-18-567131), an anonymous donor, and the Wyncote Foundation.

9. Bibliographical References

Amici, S., Ogar, J., Bozeat, S., Arnold, R., Watson, P., and Hodges, J.R. (2006). Performance in specific language tasks correlates with regional volume changes in progressive aphasia. *Cognitive and Behavioral Neurology,* 20(4): 203–211.

Ash, S., Moore, P., Vesely, L., Gunawardena, D., McMillan, C., Anderson, C., … and Grossman, M. Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics,* 22(4):370–383.

Boersma, P., and Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.09, retrieved 26 January 2020 from http://www.praat.org

Bonner, M., Price, A., Peelle, J., and Grossman, M. (2016). Semantics of the visual environment encoded in parahippocampal cortex. *Journal of Cognitive Neuroscience,* 28(3): 361–378.

Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Brysbaert, M., Mandera, P., and Keuleers, E. (2018). Word prevalence norms for 62 ,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Cho, S., Nevler, N., Ash, S., Shellikeri, S., Irwin, D., Massimo, L., Rascovsky, K., Olm, C., Grossman, M., and Liberman, M. (under review). Automated analysis of lexical features in Frontotemporal Degeneration.

Cousins, K., York, C., Bauer, L., and Grossman, M. (2016). Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia,* 84:244–251.

Fraser, K., Meltzer, J., Graham, H., Leonard, C., Hirst, G., Black, S., and Rochon, E. (2014). Automated

classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex,* 55:43–60.

Galvin, J., Howard, D., Denny, S., Dickinson, S., and Tatton, N. (2017). The social and economic burden of frontotemporal degeneration. *Neurology,* 89:2049–2056.

Goodglass, H., Kaplan, E., and Weintraub, S. (1983). Boston diagnostic aphasia examination. Philadelphia, PA, Lea & Febiger.

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology,* 76:1006–1014.

Grossman, M., Mickanin, J., Onishi, K., Hughes, E., D'Esposito, M., Ding, X. S., … and Reivich, M. (1996). Progressive nonfluent aphasia: Language, cognitive, and PET measures contrasted with probable Alzheimer's disease. *Journal of Cognitive Neuroscience*, 8(2) :135–154.

Honnibal, M., and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Màrques et al., Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1373–1378, Lisbon, Portugal, September.

Johnson, J.K., Diehl, J., Mendez, M.F., Neuhaus, J., Shapira, J.S., Forman, M., Chute, D.J., Roberson, E.D., Pace-Savitsky, C., Neumann, M., Chow T.W., Rosen, H.J., Forstl, H., Kurz, A., and Miller, B.L. Frototemporal lobar degeneration: Demographic characteristics of 353 patients. *Arch Neurol.,* 62: 925–930.

Monroe, B., Colaresi, M., and Quinn, K. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.

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

Nevler, N., Ash, S., Jester, C., Irwin D. J., Liberman, M., and Grossman, M. (2018). Automatic measurement of prosody in behavioral variant FTD. *Neurology,* 89:650–656.

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

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Gorno-Tempini, M., and Ogar, J. (2008). Learning diagnostic models using speech and language measures. In Wheeler B., et al., Proceedings of the 30th Annual International IEEE EMBS Conference, pages 4648–4651, Vancouver, Canada, August.

Ratnavalli, E., Brayne, C., Dawson, K., and Hodges, J.R. (2002). The prevalence of frontotemporal dementia. *Neurology,* 58:1615–1621.

Rhee, J., Antiquena, P., and Grossman, M. (2001). Verb comprehension in frontotemporal degeneration: The role of grammatical, semantic and executive components. *Neurocase*, 7(2):173–184.

Themistocleous, C., Ficek, B., Webster K.T., Wendt H., Hillis A.E., Den Ouden, D.B., and Tsapkini, K. Acoustic markers of PPA variants using machine learning. *Frontiers in Human Neuroscience 92.* doi: 10.3389/conf.fnhum.2018.228.00092.

DOI: 10.1002/alz.045315

BIOMARKERS

POSTER PRESENTATIONS

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

Biomarkers (non-neuroimaging) / Longitudinal change over time

A longitudinal study of automated analysis of acoustic speech markers in FTD and PPA

Naomi Nevler[1] | Sharon Ash[1] | Sunghye Cho[2] | Sanjana Shellikeri[1] | Natalia Parjane[1] | David J Irwin[1] | Mark Y Liberman[2] | Murray Grossman[1]

[1] Penn FTD Center, University of Pennsylvania, Philadelphia, PA, USA

[2] Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

Correspondence
Naomi Nevler, Penn FTD Center, University of Pennsylvania, Philadelphia, PA, USA.
Email: naomine@pennmedicine.upenn.edu

Abstract

Background: Speech is a complex activity requiring proper function and connectivity of multiple brain networks and as such is sensitive to focal neurodegeneration. We have previously reported on acoustic markers of dysprosody in speech samples of speakers with frontotemporal dementia (FTD) phenotypes. In the current study we explore the longitudinal changes in acoustic-prosodic markers in FTD.

Method: We analyzed 102 speech samples of picture descriptions from 46 participants with FTD (Table 1): 8 with non-fluent/agrammatic primary progressive aphasia (naPPA), 14 with semantic variant PPA (svPPA), 10 with logopenic aphasia (lvPPA) and 14 with behavioral FTD (bvFTD). We automatically segmented the acoustic signal into segments of continuous speech or silence, measured their durations, and derived other measures. We used linear mixed effects (lme) models to test changes over time for each acoustic measure, controlling for sex, education, and random intercepts. We also examined any interaction between phenotypes and disease duration.

Result: bvFTD speakers increased their pause duration by 0.27 seconds per year and their pause rate by 3.9 pauses per minute (ppm) each year. Their speech segment duration shortened by 0.1 seconds per year (p=0.041), decreasing their total speech time by 6.6 seconds (p=0.003) per year. Thus, bvFTD patients reduced the proportion of speech in their samples by 5.16 percent per year (p=0.008). svPPA speakers increased their pause rate similarly, but in contrast, their pause duration decreased by 0.097 seconds per year and they increased their speech segment frequency by 8.32 per minute each year (p=0.054). naPPA and lvPPA speakers increased their pause rate over time and spent less total time (speech + pause) describing the picture (by 5.6 seconds per year; p=0.018). They did not differ from bvFTD and svPPA in these two acoustic measures.

Conclusion: In our study all FTD speakers became more dysfluent and produced shorter descriptions with time, however, only bvFTD speakers actually exhibited reduced speech production. In contrast, svPPA speech had more frequent pauses and speech segments over time, rendering it "fragmented" and inefficient. These findings support the role of automated acoustic analysis in characterizing speech longitudinally in neurodegeneration.

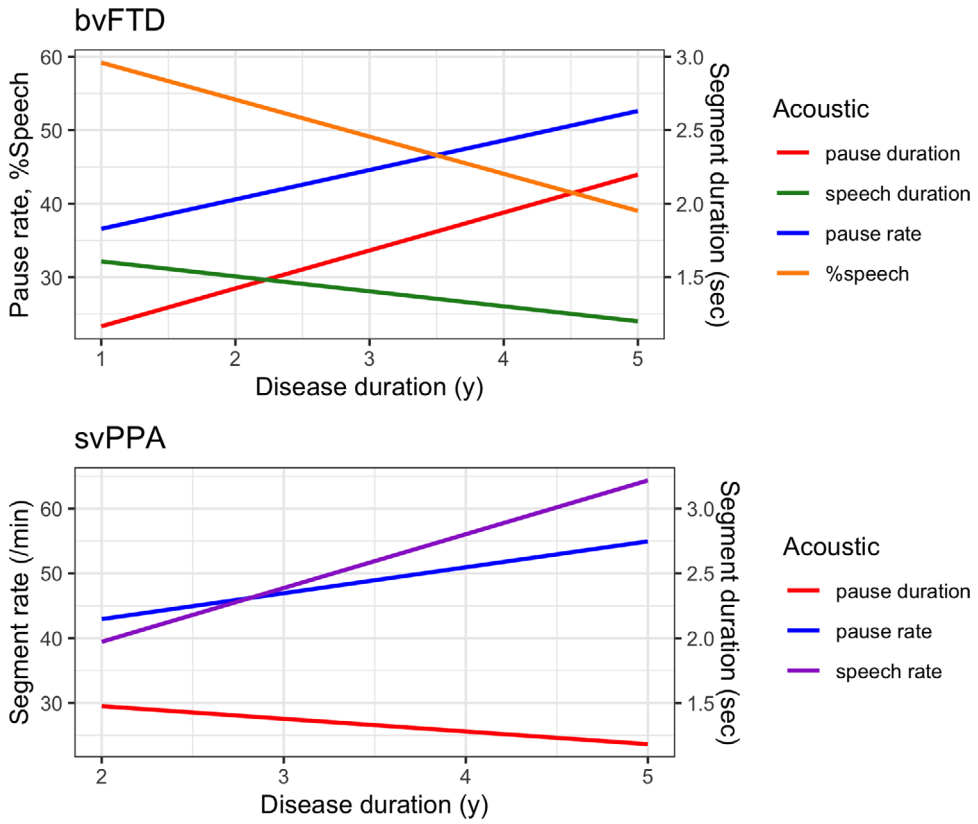


FIGURE 1

TABLE 1

Table 1: Clinical & Demographic characteristics (mean (sd))

	bvFTD	lvPPA	naPPA	svPPA	p
n	14	10	8	14	
Sex = Male (%)	12 (85.7)	3 (30.0)	0 (0.0)	5 (35.7)	0.001
Age (y)	66.29 (6.40)	63.60 (8.85)	70.62 (7.09)	61.36 (9.33)	0.074
Age at onset (y)	64.00 (5.96)	60.70 (9.19)	67.88 (6.88)	58.71 (9.45)	0.068
Education (y)	15.64 (1.82)	14.60 (2.72)	13.88 (1.73)	15.00 (2.39)	0.331
Disease duration (y)	2.29 (1.33)	2.80 (1.32)	2.75 (0.89)	2.64 (0.74)	0.656

RESEARCH ARTICLE

Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology

Sunghye Cho, PhD, Katheryn Alexandra Quilico Cousins, PhD, Sanjana Shellikeri, PhD, Sharon Ash, PhD, David John Irwin, MD, Mark Yoffe Liberman, PhD, Murray Grossman, MD, EdD, and Naomi Nevler, MD

Neurology® 2022;99:e313-e322. doi:10.1212/WNL.0000000000200581

Correspondence
Sunghye Cho
csunghye@sas.upenn.edu

Abstract

Background and Objectives

We compared digital speech and language features of patients with amnestic Alzheimer disease (aAD) or logopenic variant primary progressive aphasia (lvPPA) in a biologically confirmed cohort and related these features to neuropsychiatric test scores and CSF analytes.

Methods

We included patients with aAD or lvPPA with CSF (phosphorylated tau ([p-tau]/β-amyloid [Aβ] ≥0.09, and total tau/Aβ ≥0.34) or autopsy confirmation of AD pathology and age-matched healthy controls (HC) recruited at the Frontotemporal Degeneration Center of the University of Pennsylvania for a cross-sectional study. We extracted speech and language variables with automated lexical and acoustic pipelines from participants' oral picture descriptions. We compared the groups and correlated distinct features with clinical ratings and CSF p-tau levels.

Results

We examined patients with aAD (n = 44; age 62 ± 8 years; 24 women; Mini-Mental State Examination [MMSE] score 21.1 ± 4.8) or lvPPA (n = 21; age 64.1 ± 8.2 years; 11 women; MMSE score 23.0 ± 4.2) and HC (n = 28; age 65.9 ± 5.9 years, 15 women; MMSE score 29 ± 1). Patients with lvPPA produced fewer verbs (10.5 ± 2.3; $p = 0.001$) and adjectives (2.7 ± 1.3, $p = 0.019$) and more fillers (7.4 ± 3.9; $p = 0.022$) with lower lexical diversity (0.84 ± 0.1; $p = 0.05$) and higher pause rate (54.2 ± 19.2; $p = 0.015$) than individuals with aAD (verbs 12.5 ± 2; adjectives 3.8 ± 2; fillers 4.9 ± 4.5; lexical diversity 0.87 ± 0.1; pause rate 45.3 ± 12.8). Both groups showed some shared language impairments compared with HC. Word frequency (MMSE score: $\beta = -1.6$, $p = 0.009$; Boston Naming Test [BNT] score: $\beta = -4.36$, $p < 0.001$), adverbs (MMSE score: $\beta = -1.9$, $p = 0.003$; BNT score: $\beta = -2.41$, $p = 0.041$), pause rate (MMSE score: $\beta = -1.21$, $p = 0.041$; BNT score: $\beta = -2.09$, $p = 0.041$), and word length (MMSE score: $\beta = 1.75$, $p = 0.001$; BNT score: $\beta = 2.94$, $p = 0.003$) were significantly correlated with both MMSE and BNT scores, but other measures were not correlated with MMSE and/or BNT score. Prepositions ($r = -0.36$, $p = 0.019$), nouns ($r = -0.31$, $p = 0.047$), speech segment duration ($r = -0.33$, $p = 0.032$), word frequency ($r = 0.33$, $p = 0.036$), and pause rate ($r = 0.34$, $p = 0.026$) were correlated with patients' CSF p-tau levels.

Discussion

Our measures captured language and speech differences between the 2 phenotypes that traditional language-based clinical assessments failed to identify. This work demonstrates the potential of natural speech in reflecting underlying variants with AD pathology.

RELATED ARTICLE

Editorial
Natural Speech Analysis: A Window Into Alzheimer Disease Phenotypes
Page 137

From the Linguistic Data Consortium (S.C., M.Y.L.) and Department of Neurology (K.A.Q.C., S.S., S.A., D.J.I., M.G., N.N.), University of Pennsylvania, Philadelphia.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

Glossary

aAD = amnestic AD; **Aβ** = β-amyloid$_{42}$; **AD** = Alzheimer disease; **BNT** = Boston Naming Test; **lvPPA** = logopenic variant PPA; **p-tau** = phosphorylated tau; **POS** = part-of-speech; **PPA** = primary progressive aphasia.

Speech production is a complex behavior involving coordinated activation of multiple brain regions. Thus, examining speech production provides potential opportunities to identify neurodegenerative disease markers that are sensitive to specific phenotypes. Because Alzheimer disease (AD) accounts for up to 80% of patients with dementia,[1] much attention has been paid to cognitive and linguistic profiling of AD. Language produced by patients with amnestic AD (aAD) has been found to be "empty" with an abundance of nonspecific words, circumlocutions, and sparse content.[2]

Logopenic variant primary progressive aphasia (lvPPA) is 1 of the PPA variants[3,4] that is an atypical, nonamnestic manifestation of AD,[3,5-9] with a majority of autopsied cases associated with underlying AD pathology.[9,10] Since the identification of this PPA variant, many studies have been dedicated to characterizing its language and speech features compared to other variants of PPA.[11,12] Previous studies have shown that patients with lvPPA speak slowly with impaired lexical access[13] and have poor phonemic discrimination[14] with limited auditory-verbal short-term memory, naming impairment,[4,5] and dysfluencies.[12]

Previous studies of patients with neurodegeneration suggest that language and speech features are useful as a screening tool[15-17] because speech samples are easy to collect noninvasively and are sensitive to cognitive impairments. Despite the shared pathology of aAD and lvPPA, previous studies have focused on the linguistic profiling of these 2 syndromes separately. With few comparative studies, an important gap remains in the literature. Most previous quantitative work has focused on measures such as the Boston Naming Test (BNT) that assess lexical retrieval during confrontation naming of an object. However, descriptions of natural, connected speech in aAD and lvPPA are frequently informal. In this study, we identified similarities and differences between patients with aAD and lvPPA with biological confirmation of underlying AD pathology by analyzing digitized, natural speech samples with reliable and reproducible automated methods. From previous studies, we hypothesized that patients with lvPPA would produce more dysfluent speech with more limited lexical content than those with aAD. We also hypothesized that patients with aAD and those with lvPPA would share some linguistic features, including decreased speech production. We associated language and speech variables with clinical test scores and CSF analytes for additional validation and specific mechanistic clarification.

Methods

Standard Protocol Approvals, Registrations, and Patient Consents

The Institutional Review Board of the Hospital of the University of Pennsylvania approved the study of human participants, and all participants agreed to participate in the study by written consent. All digital data were stored in secured Health Insurance Portability and Accountability Act–compliant servers and handled by personnel trained in personal identifiable information protection.

Participants

We examined oral picture descriptions that were produced by 93 participants who were recruited in the Department of Neurology at the Hospital of the University of Pennsylvania from early 2000 to early 2019. All patients were assessed by experienced neurologists (M.G., D.J.I.) following published diagnostic criteria,[18,19] and their clinical phenotypes were reviewed in a consensus conference. To support the diagnosis, participants underwent comprehensive neuropsychological assessments with the National Alzheimer's Coordinating Center Uniform Data Set version 3[20] and the Rey Complex Figure Test.[21] Of 152 patients whose clinical phenotype was either AD (n = 114) or lvPPA (n = 38), we included only 44 participants with aAD and 21 patients with lvPPA who had AD pathology at autopsy (n = 15; 4 with lvPPA and 11 with AD) or who met the criteria of underlying AD pathology on CSF analyte levels (n = 50; 17 with lvPPA and 33 with AD; phosphorylated tau [p-tau]/β-amyloid$_{42}$ [Aβ] ≥0.09[22] and total tau/Aβ ≥0.34[23]). Five patients with aAD who had a CSF Aβ value >192 pg/mL[24] were still included in the analysis because both their p-tau/Aβ ratio and total tau/Aβ ratio met the criteria. Patients with aAD or lvPPA who did not meet both cutoffs, did not have CSF or autopsy data, or did not have AD pathology at autopsy were excluded from the analysis. Thirteen patients of 15 with autopsy data had a high probability of having AD pathologic change on the basis of established ABC scoring.[25] Four patients with posterior cortical atrophy and 3 patients with nonamnestic mild cognitive impairment who did not meet the criteria for lvPPA[18] were excluded from the analysis. None of the participants included in the study had other neurologic, psychiatric, or medical conditions that could affect cognition. Twenty-eight age-matched elderly healthy controls (HC) who did not have cognitive impairment were included as a control group.

Data Collection

We digitally recorded the participants' descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination.[26] Descriptions were a 1 minute long. Recordings

were orthographically transcribed by a linguist and trained annotators. Transcribing a 1-minute speech sample usually took 5 to 7 minutes, and our usual transcription interrater agreement rate was between 93% and 94%. The earliest recording of each participant was analyzed with our lexical and acoustic pipelines, as described below.

Lexical Pipeline

We automatically tagged the part-of-speech (POS) category of all tokens using spaCy,[27] which is a Python package for natural language processing, with its large language model (en_core_web_lg) for English. The number of tense-inflected verbs was calculated by summing the number of modal auxiliaries and past-tense and present-tense verbs. Dysfluency markers, including fillers, repetitions, and partial words, were counted separately. The count of each POS category, tense-inflected verbs, and dysfluency markers was converted to counts per 100 words to control for the total number of words per participant.

We rated each word for concreteness,[28] semantic ambiguity,[29] frequency,[30] age at acquisition,[31] and familiarity[31] using published norms. Word length by the number of phonemes for each word was calculated with the Carnegie Mellon University Pronouncing Dictionary[32] using the Natural Language Toolkit package[33] in Python. We calculated the mean scores of these measures for content words (nouns, verbs, adjectives, and adverbs) per participant.

Last, we measured lexical diversity using the moving-average type-token ratio,[34] which has been described as one of the most reliable measures for calculating lexical diversity.[35] The window length was set at 15 words. We also experimented with larger windows (20-word and 25-word windows). Because the results remained the same, we reported only results from 15-word windows. Hereafter, the measures from the lexical pipeline were referred to as language measures. Detailed descriptions of the lexical pipeline and validation of the POS tagging accuracy have previously been published.[36]

Acoustic Pipeline

We used an in-house speech activity detector to segment audio recordings into speech segments and silent pauses. We visually reviewed the segments. Nonspeech segments at the beginning and end of each recording and interviewer's prompts were excluded from the analysis.

Using the speech activity detector output, we calculated duration-related measurements, including mean speech segment duration, mean pause segment duration, percent of speech, and pause rate per minute. We summed the number of syllables of all words from the published norm[32] and computed articulation rate as the number of syllables per second.

In addition, we pitch-tracked all speech segments with Praat.[37] To normalize physiologic differences in pitch (f0), we converted the pitch values from Hertz to semitones using the 10th pitch percentile of each participant as a baseline: semitones = $12 \times \log_2(\text{f0}/\text{baseline f0})$. We used the converted 90th percentile as a measure of the pitch range of each speaker. Hereafter, the measures from the acoustic pipeline were referred to as speech measures. Detailed descriptions of the acoustic pipeline have been published previously,[38] and the list of all analyzed features is included in eTable 1, links.lww.com/WNL/B1000.

CSF Analysis

Forty-two (30 with aAD and 12 with lvPPA) patients had CSF biomarkers collected within 1 year of the Cookie Theft recording, including Aβ and p-tau. CSF was analyzed with 2 platforms, Luminex xMAP or Innotest ELISA, which was then transformed to the Luminex scale.[39] We previously related CSF p-tau levels directly to cerebral tau burden in our autopsy cohort.[39] To determine the association of language and speech features with in vivo measures of pathology, we examined the relationship between our language and speech variables and the 2 CSF biomarkers of Aβ and p-tau. Only 1 HC had CSF biomarkers in this dataset.

Statistical Methods

To compare the groups, we tested whether requirements for parametric tests were met with a Levene test. If the data met the requirements for parametric tests, we performed an analysis of variance. If not, we performed a Kruskal-Wallis test. We visually assessed residuals of the models to ensure that the data were suitable for linear modeling. When a group difference was significant, we in addition performed a post hoc pairwise *t* test or pairwise Wilcoxon rank-sum test for pairwise group comparisons, adjusting *p* values for multiple group comparisons (n = 3) with the false discovery rate. We reported the effect size of each group comparison using the Cohen *d*.

Patients' language and speech variables were *z* scored using the mean and SD of the HC. These *z* scores were used for visualization and linear regressions to estimate relations of our language variables and clinical ratings. We did not use *z* scores to determine significant group differences with *Z* tests because the test statistic in some variables did not follow a normal distribution.

The *z* scored variables that showed significant group differences were associated with patients' clinical assessments to investigate the relations of our language and speech features to clinical ratings of cognitive, language, and memory impairment. Analyzed clinical assessments included the MMSE, BNT, Rey complex figure copy and delayed recall, Craft Story delayed recall, and forward digit span scores. To examine potential interactions, we included phenotype as an interaction term (clinical ratings ~ language or speech variable × phenotype). The *p* values were adjusted with the false discovery rate.

To validate our findings with levels of specific CSF biomarkers, we correlated patients' CSF analyte levels with the language and speech variables using Pearson correlation tests. CSF p-tau levels were log-transformed to normalize the data. We also checked whether patients' clinical phenotype and the time difference between Cookie Theft recording and CSF sample collection were significant factors with linear regression models. Because the 2 factors were not significant, we reported only the results of simple correlations to simplify the models. All statistical analyses were performed with R version 4.1.0 and RStudio version 1.4.1717 (R Core Team, Vienna, Austria).

Data Availability

Anonymized data will be shared on request from any qualified investigator for purposes of validation or replication of study methods.

Results

Participant Characteristics

Table 1 shows the demographic and clinical characteristics of the participants. The 3 groups did not differ in age, sex, or education level. The patient groups did not differ from each other in disease duration, CSF biomarkers, and most clinical ratings except the Rey complex figure and forward digit span scores. Patients with aAD were more impaired in both Rey complex figure copy and delayed recall because these patients had the amnestic variant of AD. On the other hand, patients with lvPPA were more impaired on forward digit span, and this was in line with our previous observation.[9]

Differences Between Patients With aAD and Patients With lvPPA

The group means of all language and speech variables are summarized in eTable 2, links.lww.com/WNL/B1000. Patients with lvPPA produced fewer tense-inflected verbs compared to those with aAD ($p = 0.001$, $|d| = 0.94$) and HC ($p = 0.048$, $|d| = 0.61$; Figure1A). The tense-inflected verb counts of patients with aAD did not significantly differ from those of HC ($p = 0.124$, $|d| = 0.4$). Patients with lvPPA showed lower lexical diversity than those with aAD ($p = 0.05$, $|d| = 0.52$) and HC ($p = 0.005$, $|d| = 1.02$), yet patients with aAD did not differ from HC ($p = 0.149$, $|d| = 0.39$; Figure1A). Larger windows yielded similar results (20-word window: lvPPA vs aAD $p = 0.047$, $|d| = 0.53$ and lvPPA vs HC $p = 0.004$, $|d| = 1.02$; 25-word window: lvPPA vs aAD $p = 0.052$, $|d| = 0.52$ and lvPPA vs HC $p = 0.004$, $|d| = 1$). Patients with lvPPA produced fewer adjectives than those with aAD ($p = 0.019$, $|d| = 0.66$) and HC ($p < 0.001$, $|d| = 1.72$); Patients with aAD also produced fewer adjectives than HC ($p = 0.003$, $|d| = 0.75$; Figure 1A). Thus, both patient groups were impaired in their adjective production, but patients with lvPPA were more severely impaired compared with those with aAD.

Patients with lvPPA also produced more fillers than patients with aAD ($p = 0.022$, $|d| = 0.6$) and HC ($p = 0.01$, $|d| = 1.06$; Figure 1B), while patients with aAD did not differ significantly from HC ($p = 0.383$, $|d| = 0.23$). Patients with lvPPA showed a higher pause rate than those with aAD ($p = 0.015$, $|d| = 0.55$) and HC ($p < 0.001$, $|d| = 1.58$; Figure 1B). The pause rate of patients with aAD was also higher than that of HC ($p < 0.001$, $|d| = 1.34$). Last, patients with lvPPA produced more partial words than HC ($p = 0.042$, $|d| = 0.61$), but they did not differ significantly from those with aAD ($p = 0.147$, $|d| = 0.3$), and patients with aAD did not differ from HC ($p = 0.313$, $|d| = 0.24$). Thus, patients with lvPPA produced an abnormal number of partial words, while speakers with aAD did not.

Impaired Language and Speech Features in Both aAD and lvPPA

Figure 2 shows all language and speech features in which both patient groups differed from HC. Both patient groups produced fewer prepositions and nouns than HC; patients produced shorter speech segments than HC, and their percent of speech time of the total time was also lower than that of HC ($p < 0.001$, $|d| > 0.8$ for all comparisons). Both groups' content words were shorter, more frequent (length and frequency: $p < 0.001$, $|d| > 0.8$), acquired earlier (lvPPA: $p < 0.001$, $|d| = 0.95$; aAD: $p < 0.001$, $|d| = 0.56$), and more concrete (lvPPA: $p = 0.028$, $|d| = 0.7$; aAD: $p = 0.002$, $|d| = 0.86$) than those of HC. Both patient groups produced more adverbs (lvPPA: $p = 0.029$, $|d| = 0.84$; aAD: $p = 0.029$, $|d| = 0.73$) and repetitions (lvPPA: $p < 0.001$, $|d| = 1.39$; aAD: $p < 0.001$, $|d| = 0.79$) than HC. Patients also spoke more slowly (lvPPA: $p < 0.001$, $|d| = 1.01$; aAD: $p < 0.001$, $|d| = 0.57$), and they produced fewer words in total than HC (lvPPA: $p = 0.032$, $|d| = 0.7$; aAD: $p = 0.007$, $|d| = 0.73$).

Relationships to Clinical Measures

Figure 3 illustrates relationships between the speech and language features and 2 clinical ratings: MMSE and BNT. Only 1 variable showed a significant interaction with phenotype. Patients with aAD who had lower MMSE scores produced more adverbs ($\beta = -1.9$, $p = 0.003$), yet patients with lvPPA who had lower MMSE scores produced fewer adverbs ($\beta = 2.67$, $p = 0.015$).

MMSE score was significantly related to 5 language and speech variables. Patients with lower MMSE scores produced more frequent ($\beta = -1.6$, $p = 0.009$) and shorter ($\beta = 1.75$, $p = 0.001$) content words, paused more frequently ($\beta = -1.21$, $p = 0.041$), and produced more adverbs ($\beta = -1.9$, $p = 0.003$) and partial words ($\beta = -1.52$, $p = 0.021$).

BNT score was significantly associated with 10 variables. Patients with lower BNT scores produced frequent words ($\beta = -4.36$, $p < 0.001$), produced many adverbs ($\beta = -2.41$, $p = 0.041$), and had a high pause rate ($\beta = -2.09$, $p = 0.041$). Patients with lower BNT scores also produced fewer prepositions ($\beta = 2.42$, $p = 0.048$) and nouns ($\beta = 5.18$, $p = 0.003$). Patients with lower BNT scores showed a low percent of speech produced during the picture description ($\beta = 2.4$, $p =$

Table 1 Demographic and Clinical Characteristics of Participants

	aAD (n = 44)	lvPPA (n = 21)	*p* Value	HC (n = 28)	*p* Value
Age, y	62.0 (8.0)	64.1 (8.2)	0.335	65.9 (5.9)	0.098
Education, y	16.0 (2.6)	16.1 (3.2)	0.919	15.9 (2.6)	0.965
Sex (male), n (%)	20 (45.5)	10 (47.6)	0.870	13 (46.4)	0.986
Disease duration, y	3.7 (2.6)	3.4 (1.5)	0.599	NA	
CSF p-tau	38.5 (20.1)	38.4 (16.0)	0.982	12.0 (NA)	0.387
CSF Aβ	139.9 (36.2)	135.0 (26.2)	0.605	415.0 (NA)	<0.001
CSF p-tau/Aβ	0.3 (0.2)	0.3 (0.2)	0.877	0.0 (NA)	0.297
MMSE score, n	44	21	0.133	28	<0.001
Mean (SD)	21.1 (4.8)	23.0 (4.2)		29.2 (1.0)	
BNT score, n	42	19	0.513	21	<0.001
Mean (SD)	20.1 (8.4)	18.6 (8.6)		27.9 (2.7)	
Animal fluency score, n	36	19	0.856	21	<0.001
Mean (SD)	10.6 (4.9)	10.4 (4.2)		20.6 (6.0)	
F-letter fluency score, n	25	6	0.141	1	0.049
Mean (SD)	8.8 (3.8)	6.3 (2.4)		16.0 (NA)	
Digit span forward score, n	31	15	<0.001	6	<0.001
Mean (SD)	5.8 (2.1)	3.2 (1.7)		9.0 (1.8)	
Rey figure copy score, n	29	14	0.005	21	<0.001
Mean (SD)	18.8 (13.4)	30.1 (6.2)		34.7 (2.1)	
Rey figure delayed score, n	28	14	0.010	21	<0.001
Mean (SD)	6.9 (7.5)	13.6 (7.5)		19.0 (7.2)	
Craft story delayed score, n	19	6	0.898	1	0.005
Mean (SD)	8.9 (9.8)	9.5 (5.8)		43.0 (NA)	

Abbreviations: aAD = amnestic Alzheimer disease; Aβ = β-amyloid$_{42}$; BNT = Boston Naming Test; HC = healthy controls; lvPPA = logopenic variant primary progressive aphasia; MMSE = Mini-Mental State Examination; NA = not applicable; p-tau = phosphorylated tau.
The patient groups were compared with *t* tests except for the sex ratio. The comparisons of all groups were tested with analysis of variance except for the sex ratio. The sex ratios were compared with χ^2 tests.

0.045), and they produced shorter (β = 2.94, $p = 0.003$), less concrete (β = 4.24, $p = 0.003$), and earlier-acquired (β = 4.69, $p = 0.003$) content words with shorter speech segments (β = 3.89, $p = 0.041$).

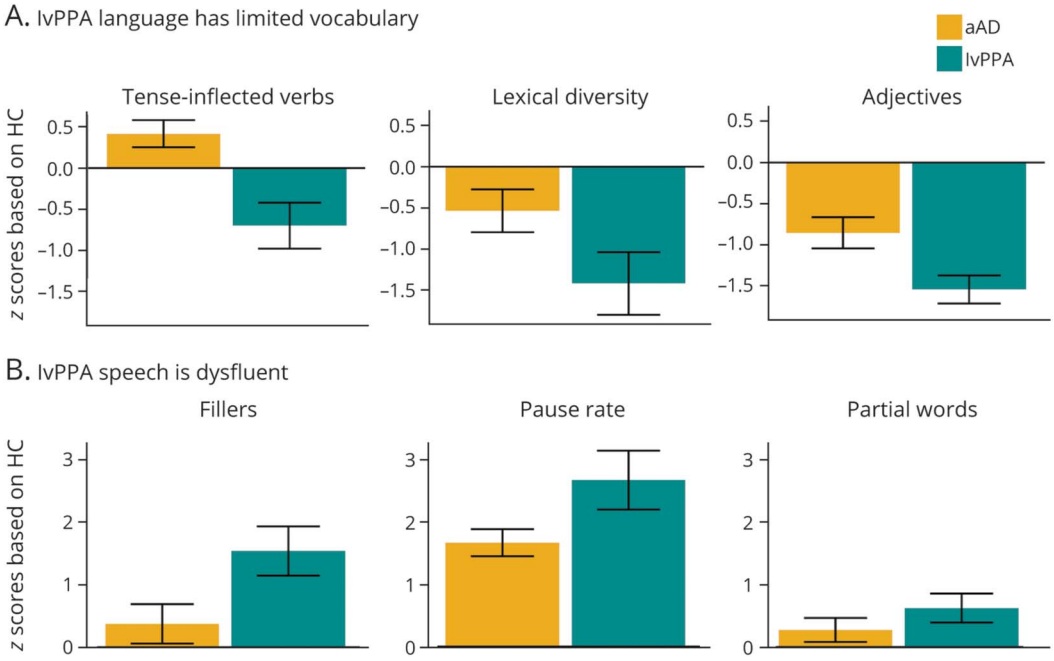
The other clinical ratings, namely the Rey complex figure copy and delayed recall, the Craft Story delayed recall, and the forward digit span, and their interaction with patients' phenotype were not significantly related to our language/speech measures after *p* value adjustments for multiple comparisons.

CSF Results

Forty-two (30 with aAD and 12 with lvPPA) patients had CSF biomarkers, including p-tau, collected within 1 year of the Cookie Theft recording (mean interval 4.7 ± 3.5 months). This subset did not differ demographically or clinically from the larger group of patients. The aAD and lvPPA subset groups did not differ in age ($p = 0.64$), sex ($p = 0.99$), education ($p = 0.82$), disease duration ($p = 0.75$), or time difference between CSF sample collection and the Cookie Theft recording ($p = 0.43$).

Patients' CSF p-tau level correlated with lower preposition counts ($r = -0.36$, $p = 0.019$), lower noun counts ($r = -0.31$, $p = 0.047$), and higher word frequency ($r = 0.33$, $p = 0.036$; Figure 4A). In addition, patients' p-tau levels were correlated with a higher pause rate ($r = 0.34$, $p = 0.026$) and shorter mean speech segment durations ($r = -0.33$, $p = 0.032$; Figure 4B). The other measures (adverbs, partial words, tense-inflected verbs, articulation rate, lexical diversity, total words, fillers, repetitions, adjectives, percent of speech, word length, concreteness, age at acquisition) were not related to patients' p-tau levels. Aβ alone did not significantly correlate with any of the language measures.

Figure 1 Speech Differences Between lvPPA and aAD

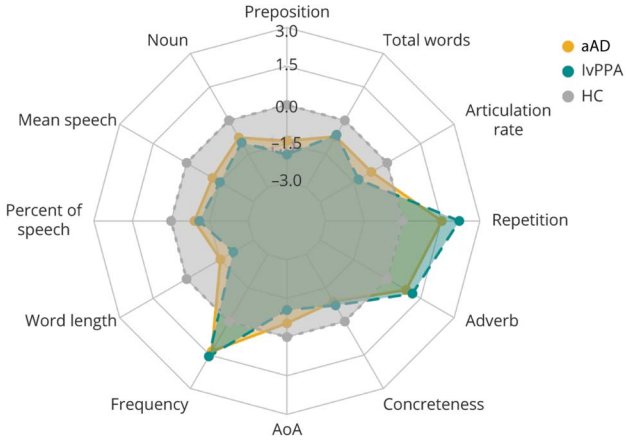


For ease of visualization, *z*-scored values compared to HC mean and standard deviation are plotted (A and B). Horizontal black line indicates the mean of the healthy controls (HC). aAD = amnestic Alzheimer disease, lvPPA = logopenic variant primary progressive aphasia.

Discussion

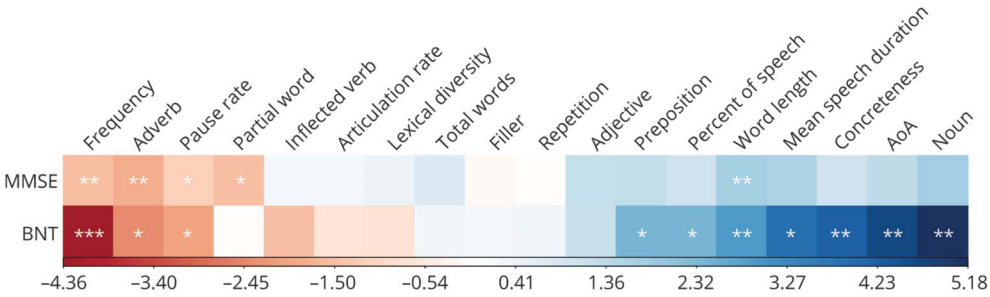
LvPPA is most frequently associated with underlying AD pathology, but direct comparison of lvPPA with aAD has been reported rarely. Patients with lvPPA have been compared to patients with the other types of PPA, who generally have frontotemporal lobar degeneration pathology.[11,12] While it may be clear that patients with lvPPA differ from patients with the other PPA variant due to different pathology, it remains to be seen how patients with nonamnestic lvPPA differ from patients with aAD who have the same pathology. To fill this gap in the literature, the current study focuses on characterizing the language similarities and differences between lvPPA and aAD in a biologically confirmed cohort. To optimize reliability and reproducibility, we used fully automated lexical and acoustic analyses to characterize language and speech markers of AD pathology. We expected that patients with nonamnestic lvPPA would produce more dysfluent speech and have limited vocabulary compared to patients with aAD because of the phenotypic characteristics of lvPPA. Results confirmed that patients with lvPPA produced fewer adjectives and tense-inflected verbs with lower lexical diversity than patients with aAD and HC. Patients with lvPPA also paused more frequently and produced more fillers and partial words than HC and patients with aAD. It is important to note that the patient groups did not differ in brief language-based clinical assessments such as animal and letter fluency tasks or the Craft Story delayed recall test, even though they differed on some of our language and speech measures, highlighting the importance of monitoring the language and speech characteristics of these patients. We also found that both patient groups shared impairments in some language and speech features relative to HC. For example, both patient groups produced more adverbs, including words like "there" and "here," but fewer prepositions and nouns than HC. In addition, patients' content words were acquired earlier, shorter, more frequent, and less concrete than those of HC.

Figure 2 Impaired Language and Speech Measures in Both lvPPA (Green) and aAD (Yellow)



Only speech measures that were significantly different from those of healthy controls (HC) (gray) are plotted, and measures were standardized on the basis of the HC mean and SD. Numbers in blue indicate *z*-scored values based on HC. aAD = amnestic Alzheimer disease; AoA = age of acquisition; lvPPA = logopenic variant primary progressive aphasia.

Figure 3 Results of Linear Regression Models of Language and Speech Measures in Patients With MMSE and BNT Scores



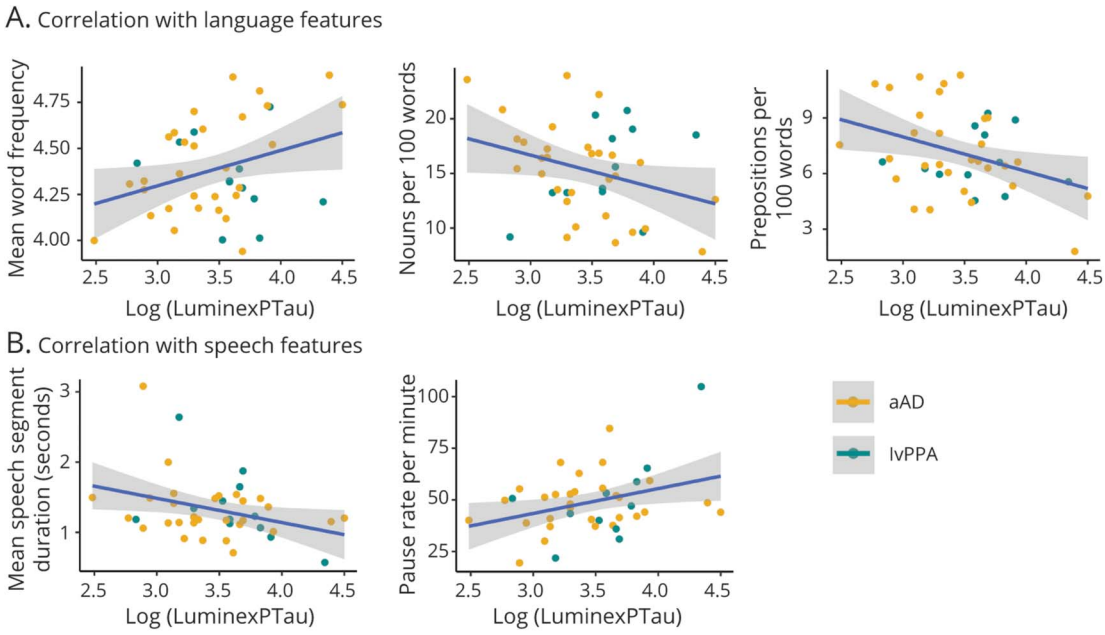
Standardized estimated coefficients (x-axis) are plotted with colors: positive coefficients in blue and negative coefficients in red. The *p* values are adjusted with the false discovery rate. AoA = age of acquisition; BNT = Boston Naming Test; MMSE = Mini-Mental State Examination. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

Patients produced more repetitions and fewer total words with a slower articulation rate and a shorter mean speech duration than HC. Some of these variables were significantly related to clinical test scores and p-tau levels in CSF. We discuss these findings below.

The patient groups significantly differed on 6 language and speech measures: pause rate, partial words, fillers, adjectives, tense-inflected verbs, and lexical diversity. Fillers, adjectives, tense-inflected verbs, and lexical diversity were significantly more impaired in patients with lvPPA than those with aAD, emphasizing the deficits in lexical retrieval and poor fluency in these patients. However, these language and speech features distinguishing between lvPPA and aAD were not related to any of the clinical ratings that we examined, including MMSE, BNT, the Rey complex figure, the Craft Story delayed recall, and forward digit span. It is important to note that the mechanism thought to subserve retrieval of a single word in response to a stimulus picture or recall of episodic memory appears to differ from lexical retrieval during natural connected speech.[40] It is thus critical to monitor these speech features because they are not easily explained by more general and commonly used clinical measures.

The result that tense-inflected verb production differed between lvPPA and aAD has rarely been reported. This finding seems to suggest that patients with lvPPA produced fewer complete sentences, assuming that there was at least 1 tense-inflected verb per inflection phrase. This would be consistent with the observation that patients with lvPPA who had limited auditory-verbal short-term memory produced briefer sentences.[3,4] Frequent fillers in lvPPA were in line with previous observations and consistent with their limited lexical retrieval.[12,13,18] Lower adjective counts in patients with aAD

Figure 4 Significant Correlations of the Speech Variables and CSF p-tau Levels (A and B)



aAD = amnestic Alzheimer disease; lvPPA = logopenic variant primary progressive aphasia; P-tau = phosphorylated tau.

compared to HC have been previously reported,[41] yet we further showed that adjective production was more impaired in nonamnestic lvPPA than in aAD. Lexical diversity has frequently been examined in the AD literature,[15,16,42] and previous studies have found that the lexical diversity of patients with aAD was lower than that of HC. We showed that lexical diversity was even lower in lvPPA than in aAD. The fact that patients with lvPPA and aAD significantly differed on these measures suggests that our language and speech variables may capture subtle but unique phenotypic differences between lvPPA and aAD. In addition, traditional clinical ratings are relatively insensitive to these linguistic features. In addition, none of the 6 variables except pause rate correlated with CSF p-tau. This may suggest that these markers are related more narrowly to the phenotype. Further studies, including the anatomic distribution of pathology in an autopsy cohort with quantitative measures of pathologic burden, may help shed light on this issue.

Some language and speech measures were related to MMSE and BNT scores but not to the other clinical ratings, including the Craft Story delayed recall, Rey figure copy and delayed recall, and forward digit span. Because the Rey figure copy is not worse in lvPPA than aAD, a visual-perceptual deficit leading to difficulty perceiving the stimulus picture is unlikely to account for the observed distinct speech and language deficits in lvPPA. Pause rate showed more impairment in lvPPA than aAD. This might indicate word-finding difficulty in lvPPA that could provoke frequent pausing to recall an appropriate word from the lexicon. It could also be that patients with lvPPA spoke slowly—these patients' articulation rate was lower than that of HC—due to their difficulty in retrieving words to generate utterances. Pause rate was significantly related to both MMSE score, an indicator of general cognitive impairment, and BNT score, a measure of confrontation naming. Elevated pause rate therefore may reflect in part both patients' word-finding difficulties and their general cognitive impairments. Partial word count, which was impaired only in lvPPA, correlated only with MMSE score but not BNT score, suggesting that it reflected in part disease severity and general cognitive impairments of patients with lvPPA but not impaired object naming.

Word-finding difficulty in aAD and lvPPA has previously been noted[13,43-47]; studies have shown that patients had impairments in auditory-verbal short-term memory and could not recall the phonologic form of a word. However, comparative studies have not been reported to examine whether both patients with aAD and patients with nonamnestic lvPPA would show word-finding difficulty to a similar degree during natural speech. In our study, both patient groups produced content words that were more abstract, acquired earlier, more frequent and shorter than those of HC, suggesting that they had difficulties in retrieving the full spectrum of lexical items needed to describe the picture. Word frequency and length were significantly related to both MMSE and BNT scores, which suggests that these lexical measures reflect in part patients' disease severity and difficulties in lexical retrieval during confrontation naming. On the other hand, concreteness and age at acquisition were significantly associated only with BNT score, indicating that these may be more sensitive to word-finding difficulty in patients. Patients with lvPPA and aAD did not significantly differ on these measures, confirming that some degree of word-finding difficulty is present in both aAD and nonamnestic lvPPA.

Adverb counts were greater in patients compared to HC. This may be related to patients frequently using proadverbs, including "here" and "there," which replaced locational prepositional phrases. Patients typically produced utterances like "Mom is standing here," for example, when HC produced "Mom is standing in front of the sink." Elevated adverb counts were associated with low BNT scores, suggesting that greater adverb use reflected patients' difficulties in naming specific locations during natural speech. Patients' difficulty in producing locational phrases was also partly reflected in the decreased preposition counts compared to HC, which was also related to BNT scores. On the contrary, patients' pronoun counts did not differ from those of HC; thus, increased adverbs and decreased prepositions seem to support the inference that patients had relatively more difficulty naming locations than naming objects. Additional work is needed to determine whether these language markers are related to temporal propositions and other features in connected speech.

Some of our language and speech variables correlated with CSF p-tau levels but not with Aβ. This finding is in line with previous findings that patients' cognitive impairment is generally not related to Aβ levels but to accumulation of p-tau.[48-50] Speech production is one of the most essential daily functions of humans, which needs to be taken into consideration in AD clinical trials and may serve in monitoring response to treatment. Because our automated procedures for collecting speech features are highly reliable and reproducible, investigations of speech variables as secondary outcome measures should be considered in disease-modifying trials targeting tau.

It is a strength of our study that we directly compared language and speech features in patients with aAD and lvPPA with biological evidence of underlying AD pathology. We examined natural connected speech quantitatively using automated analyses of digitized speech samples. We inspected differences and similarities among these groups and showed that our variables could capture subtle linguistic differences between the 2 phenotypes, and traditional cognitive measures appear to be insensitive to some of the features that distinguish aAD and lvPPA. These methods may be useful in monitoring disease progression and response to therapeutic interventions because collecting 1-minute speech samples is easy, highly reproducible, and inexpensive and can be done remotely compared to collection of other biomarkers. Our ongoing projects are currently testing the value of these

language and speech features in longitudinal datasets and developing machine learning classifiers for distinguishing patients with AD pathology from those with other types of neurodegenerative changes such as frontotemporal lobar degeneration.

Limitations of this study include the assessment of relatively small samples, the difficulty obtaining train-test generalizability data in rare lvPPA and early-onset aAD cases such as these, the use of a single stimulus picture to elicit the speech sample, and the absence of high-resolution MRI data to assess the anatomic associates of these linguistic features. In addition, we had only 12 patients with lvPPA with CSF biomarkers within 1 year of the picture description data collection, so we were not able to examine the relations between CSF biomarkers and the language/speech variables for each group. Future study with more CSF data will be needed to explore each group's relations between language/speech measures and CSF biomarkers. Last, because we had only 1 HC with CSF data, we were not able to determine whether our language and speech measures are able to distinguish HC with positive CSF AD biomarkers from those with negative CSF AD biomarkers. The relation between CSF biomarkers and language/speech variables in HC will need to be studied further in future research.

We implemented automated methods to analyze acoustic and lexical characteristics of the natural speech of patients with aAD and lvPPA. We identified language and speech markers that differed between the groups. We also found language and speech markers that were shared between these 2 AD phenotypes. This work demonstrates the potential of natural speech to reflect underlying AD pathology while distinguishing between specific phenotypes with the same pathology. Considering the cost-effectiveness and reliability of speech data, such markers could contribute to monitoring of patients for AD clinical trials in a more precise and inclusive way.

Acknowledgment

The authors thank the patients and caregivers who participated in this study.

Study Funding

This study was funded by grants from the NIH (AG066597, AG054519, NS109260, P30 AG072979, AG073510-01), Alzheimer's Association (AACSF-18-567131, AARF-D-619473, AARF-D-619473-RAPID, AARF-21-851126), and the Department of Defense (W81XWH-20-1-0531).

Disclosure

M. Grossman participates in clinical trials sponsored by Alector, Eisai, and Biogen that are unrelated to this study. He also receives research support from Biogen and Avid that is unrelated to this study and research support from NIH. M. Liberman serves on the Scientific Advisory Board for Baidu Research, USA, and is a coeditor of the *Annual Review of Linguistics*. All other authors (S. Cho, K.A.Q. Cousins, S. Shellikeri, S. Ash, N. Nevler, D.J. Irwin) report no disclosures relevant to the manuscript. Go to Neurology.org/N for full disclosures.

Publication History

Previously published at medRxiv: https://doi.org/10.1101/2021.09.27.21264148. Received by *Neurology* November 16, 2021. Accepted in final form March 8, 2022. Submitted and externally peer reviewed. The handling editor was Linda Hershey, MD, PhD, FAAN.

Appendix Authors

Name	Location	Contribution
Sunghye Cho, PhD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; study concept or design; analysis or interpretation of data
Katheryn Alexandra Quilico Cousins, PhD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; study concept or design; analysis or interpretation of data
Sanjana Shellikeri, PhD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
Sharon Ash, PhD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
David John Irwin, MD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data
Mark Yoffe Liberman, PhD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; analysis or interpretation of data
Murray Grossman, MD, EdD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data
Naomi Nevler, MD	University of Pennsylvania, Philadelphia	Drafting/revision of the manuscript for content, including medical writing for content; study concept or design; analysis or interpretation of data

References

1. Alzheimer's Association. 2020 Alzheimer's Disease Facts and Figures. 2020. Accessed May 9, 2022. https://www.alz.org/media/documents/alzheimers-facts-and-figures.pdf
2. Nicholas M, Obler L, Albert M, Goodglass H. Lexical retrieval in healthy aging. *Cortex*. 1985;21(4):595-606. doi:10.1016/S0010-9452(58)80007-6
3. Henry ML, Gorno-Tempini ML. The logopenic variant of primary progressive aphasia. *Curr Opin Neurol*. 2010;23(6):633-637. doi:10.1097/WCO.0b013e32833fb93e
4. Gorno-Tempini ML, Brambati SM, Ginex V, et al. The logopenic/phonological variant of primary progressive aphasia. *Neurology*. 2008;71:1227-1234.
5. Gorno-Tempini ML, Dronkers NF, Rankin KP, et al. Cognition and anatomy in three variants of primary progressive aphasia. *Ann Neurol*. 2004;55:335-346.
6. Rohrer JD, Rossor MN, Warren JD. Alzheimer's pathology in primary progressive aphasia. *Neurobiol Aging*. 2012;33(4):744-752. doi:10.1016/j.neurobiolaging.2010.05.020
7. Gefen T, Gasho K, Rademaker A, et al. Clinically concordant variations of Alzheimer pathology in aphasic versus amnestic dementia. *Brain*. 2012;135(5):1554-1565. doi:10.1093/brain/aws076

8. Leyton CE, Britton AK, Hodges JR, Halliday GM, Kril JJ. Distinctive pathological mechanisms involved in primary progressive aphasias. *Neurobiol Aging*. 2016;38:82-92. doi:10.1016/j.neurobiolaging.2015.10.017
9. Giannini LAA, Irwin DJ, Mcmillan CT, et al. Clinical marker for Alzheimer disease pathology in logopenic primary progressive aphasia. *Neurology*. 2017;88(24):2276-2284. doi:10.1212/WNL.0000000000004034
10. Bergeron D, Gorno-Tempini ML, Rabinovici GD, et al. Prevalence of amyloid-β pathology in distinct variants of primary progressive aphasia. *Ann Neurol*. 2018;84(5):729-740. doi:10.1002/ana.25333
11. Nevler N, Ash S, Irwin DJ, Liberman M, Grossman M. Validated automatic speech biomarkers in primary progressive aphasia. *Ann Clin Transl Neurol*. 2019;6(1):4-14. doi:10.1002/acn3.653
12. Ash S, Evans E, O'Shea J, et al. Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology*. 2013;81(4):329-336. doi:10.1212/WNL.0b013e31829c5d0e
13. Wilson SM, Henry ML, Besbris M, et al. Connected speech production in three variants of primary progressive aphasia. *Brain*. 2010;133:2069-2088. doi:10.1093/brain/awq129
14. Johnson JCS, Jiang J, Bond RL, et al. Impaired phonemic discrimination in logopenic variant primary progressive aphasia. *Ann Clin Transl Neurol*. 2020:1-6. doi:10.1002/acn3.51101
15. Fraser K, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis*. 2016;49(2):407-422. doi:10.3233/JAD-150520
16. 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(1):71-91. doi:10.1080/026870300401603
17. Rentoumi V, Paliouras G, Danasi E, et al. Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: a computational linguistics analysis. In: 8th IEEE International Conference on Cognitive Infocommunications. 2017;2017):33-38. doi:10.1109/CogInfoCom.2017.8268212
18. Gorno-Tempini ML, Hillis A, Weintraub S, et al. Classification of primary progressive aphasia and its variants. *Neurology*. 2011;76(11):1006-1014.
19. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269. doi:10.1016/j.jalz.2011.03.005
20. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Dis Assoc Disord*. 2018;32(4):351-358. doi:10.1097/WAD.0000000000000279
21. Lezak M, Howieson DB, Loring DW. *Neuropsychological Assessment*. Oxford University Press; 1983.
22. Irwin D, McMillan CT, Toledo JB, et al. Comparison of cerebrospinal fluid levels of tau and Aβ 1-42 in Alzheimer disease and frontotemporal degeneration using 2 analytical platforms. *Arch Neurol*. 2012;69(8):1018-1025. doi:10.1001/archneurol.2012.26
23. Lleó A, Irwin DJ, Illán-Gala I, et al. A 2-step cerebrospinal algorithm for the selection of frontotemporal lobar degeneration subtypes. *JAMA Neurol*. 2018;75(6):738-745. doi:10.1001/jamaneurol.2018.0118
24. Cousins KAQ, Irwin DJ, Wolk DA, et al. ATN status in amnestic and non-amnestic Alzheimer's disease and frontotemporal lobar degeneration. *Brain*. 2020;143(7):2295-2311. doi:10.1093/brain/awaa165
25. Montine TJ, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol*. 2012;123(1):1-11. doi:10.1007/s00401-011-0910-3
26. Goodglass H, Kaplan E, Weintraub S. *Boston Diagnostic Aphasia Examination*. Lea & Febiger; 1983.
27. Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. In: EMNLP 2015: Conference on Empirical Methods in Natural Language Processing; 2015:1373-1378. doi:10.18653/v1/d15-1162
28. Brysbaert M, Warriner AB, Kuperman V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods*. 2014;46(3):904-911. doi:10.3758/s13428-013-0403-5
29. Hoffman P, Lambon Ralph MA, Rogers TT. Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behav Res Methods*. 2013;45(3):718-730. doi:10.3758/s13428-012-0278-x
30. Brysbaert M, New B. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods*. 2009;41(4):977-990. doi:10.3758/BRM.41.4.977
31. Brysbaert M, Mandera P, Keuleers E. Word prevalence norms for 62,000 English lemmas. *Behav Res Methods*. 2019;51(2):467-479.
32. Carnegie Mellon Speech Group. The Carnegie Mellon University Pronouncing Dictionary. 2014. Accessed November 16, 2019. speech.cs.cmu.edu/cgi-bin/cmudict.
33. Loper E, Bird S. NLTK: the Natural Language Toolkit. In: *ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics. 2002;63-70.
34. Covington MA, McFall JD. Cutting the gordian knot: the moving average type-token ratio (MATTR). *J Quant Linguist*. 2010;17(2):94-100.
35. Fergadiotis G, Wright HH, Green SB. Psychometric evaluation of lexical diversity indices: assessing length effects. *J Speech Lang Hear Res*. 2015;58:840-852. doi:10.1044/2015
36. Cho S, Nevler N, Ash S, et al. Automated analysis of lexical features in frontotemporal degeneration. *Cortex*. 2021;137:215-231. doi:10.1016/j.cortex.2021.01.012
37. Boersma P, Weenink D. Praat: Doing Phonetics by Computer. 2019. Accessed December 3, 2021. https://www.fon.hum.uva.nl/praat/
38. Nevler N, Ash S, Jester C, Irwin DJ, Liberman M, Grossman M. Automatic measurement of prosody in behavioral variant FTD. *Neurology*. 2017;89(7):1-8.
39. Irwin DJ, Lleó A, Xie SX, et al. Ante mortem cerebrospinal fluid tau levels correlate with postmortem tau pathology in frontotemporal lobar degeneration. *Ann Neurol*. 2017;82(2):247-258. doi:10.1002/ana.24996
40. Levelt WJM. *Speaking: From Intention to Articulation*. MIT Press; 1989.
41. Croisile B, Ska B, Brabant MJ, et al. Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain Lang*. 1996;53(1):1-19. doi:10.1006/brln.1996.0033
42. Kavé G, Goral M. Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *J Clin Exp Neuropsychol*. 2016;38(9):958-966. doi:10.1080/13803395.2016.1179266
43. Buxbaum LJ, Schwartz MF, Carew TG. The role of semantic memory in object use. *Cogn Neuropsychol*. 1997;14(2):219-254. doi:10.1080/026432997381565
44. Ochipa C, Rothi LJG, Heilman KM. Conceptual apraxia in Alzheimer's disease. *Brain*. 1992;115(4):1061-1071. doi:10.1093/brain/115.4.1061
45. Adlam ALR, Bozeat S, Arnold R, Watson P, Hodges JR. Semantic knowledge in mild cognitive impairment and mild Alzheimer's disease. *Cortex*. 2006;42(5):675-684. doi:10.1016/S0010-9452(08)70404-0
46. Giffard B, Desgranges B, Nore-Mary F, et al. The nature of semantic memory deficits in Alzheimer's disease: new insights from hyperpriming effects. *Brain*. 2001;124(8):1522-1532. doi:10.1093/brain/124.8.1522
47. Venneri A, McGeown WJ, Hietanen HM, Guerrini C, Ellis AW, Shanks MF. The anatomical bases of semantic retrieval deficits in early Alzheimer's disease. *Neuropsychologia*. 2008;46(2):497-510. doi:10.1016/j.neuropsychologia.2007.08.026
48. Huber CM, Yee C, May T, Dhanala A, Mitchell CS. Cognitive decline in preclinical Alzheimer's disease: amyloid-beta versus tauopathy. *J Alzheimers Dis*. 2018;61(1):265-281. doi:10.3233/JAD-170490
49. Dickson DW, Crystal HA, Mattiace LA, et al. Identification of normal and pathological aging in prospectively studied nondemented elderly humans. *Neurobiol Aging*. 1992;13(1):179-189. doi:10.1016/0197-4580(92)90027-U
50. Engelborghs S, Maertens K, Vloeberghs E, et al. Neuropsychological and behavioural correlates of CSF biomarkers in dementia. *Neurochem Int*. 2006;48(4):286-295. doi:10.1016/j.neuint.2005.11.002