BRAIN COMMUNICATIONS

Lexical markers of disordered speech in primary progressive aphasia and 'Parkinson-plus' disorders

Shalom K. Henderson,[1,2] Siddharth Ramanan,[1] Karalyn E. Patterson,[1,2,3] Peter Garrard,[4] Nikil Patel,[4] Katie A. Peterson,[2] Ajay Halai,[1] Stefano F. Cappa,[5,6] James B. Rowe[1,2,3,†] and Matthew A. Lambon Ralph[1,†]

† These authors contributed equally to this work.

Connected speech samples elicited by a picture description task are widely used in the assessment of aphasias, but it is not clear what their interpretation should focus on. Although such samples are easy to collect, analyses of them tend to be time-consuming, inconsistently conducted and impractical for non-specialist settings. Here, we analysed connected speech samples from patients with the three variants of primary progressive aphasia (semantic, svPPA $N = 9$; logopenic, lvPPA $N = 9$; and non-fluent, nfvPPA $N = 9$), progressive supranuclear palsy (PSP Richardson's syndrome $N = 10$), corticobasal syndrome (CBS $N = 13$) and age-matched healthy controls ($N = 24$). There were three principal aims: (i) to determine the differences in quantitative language output and psycholinguistic properties of words produced by patients and controls, (ii) to identify the neural correlates of connected speech measures and (iii) to develop a simple clinical measurement tool. Using data-driven methods, we optimized a 15-word checklist for use with the Boston Diagnostic Aphasia Examination 'cookie theft' and Mini Linguistic State Examination 'beach scene' pictures and tested the predictive validity of outputs from *least absolute shrinkage and selection operator* (LASSO) models using an independent clinical sample from a second site. The total language output was significantly reduced in patients with nfvPPA, PSP and CBS relative to those with svPPA and controls. The speech of patients with lvPPA and svPPA contained a disproportionately greater number of words of both high frequency and high semantic diversity. Results from our exploratory voxel-based morphometry analyses across the whole group revealed correlations between grey matter volume in (i) bilateral frontal lobes with overall language output, (ii) the left frontal and superior temporal regions with speech complexity, (iii) bilateral frontotemporal regions with phonology and (iv) bilateral cingulate and subcortical regions with age of acquisition. With the 15-word checklists, the LASSO models showed excellent accuracy for within-sample *k*-fold classification (over 93%) and out-of-sample validation (over 90%) between patients and controls. Between the motor disorders (nfvPPA, PSP and CBS) and lexico-semantic groups (svPPA and lvPPA), the LASSO models showed excellent accuracy for within-sample *k*-fold classification (88–92%) and moderately good (59–74%) differentiation for out-of-sample validation. In conclusion, we propose that a simple 15-word checklist provides a suitable screening test to identify people with progressive aphasia, while further specialist assessment is needed to differentiate accurately some groups (e.g. svPPA versus lvPPA and PSP versus nfvPPA).

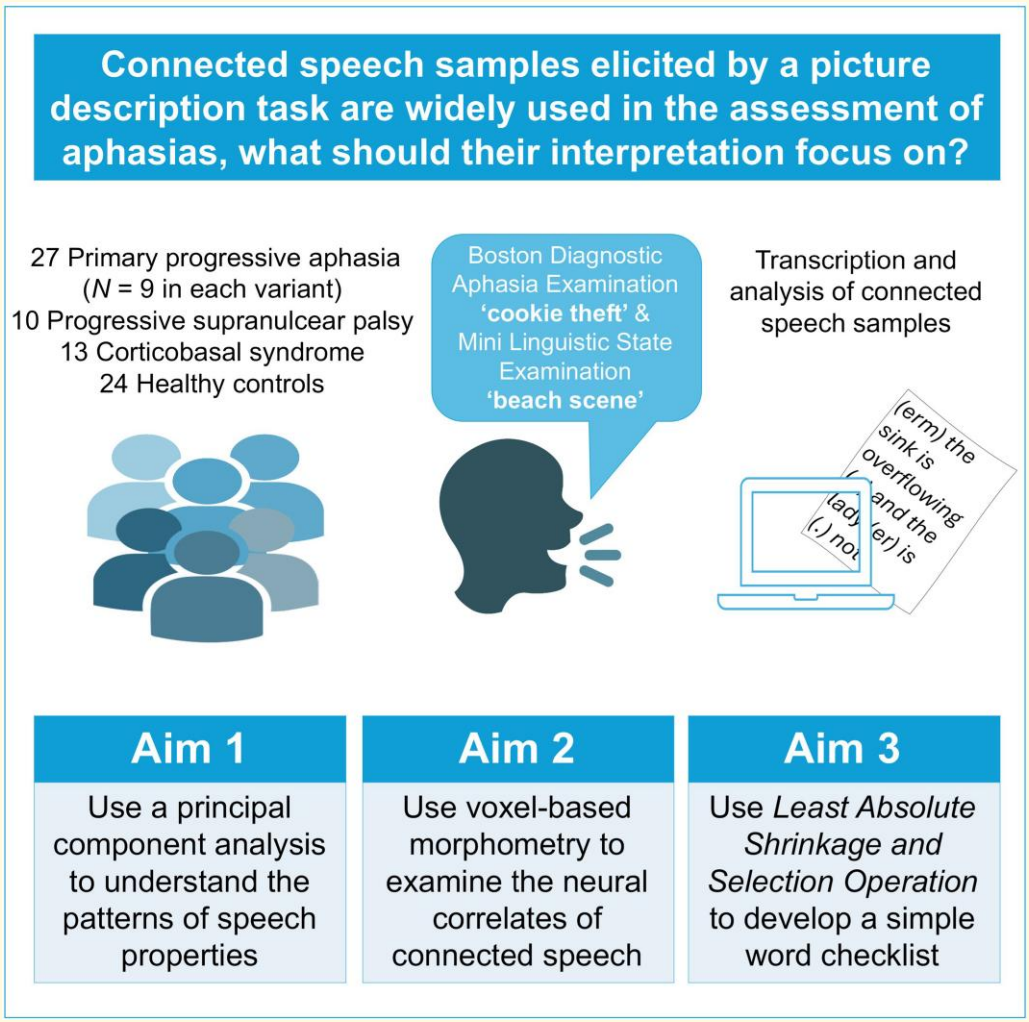
1 Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 7EF, UK
2 Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 0QQ, UK
3 Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK
4 Molecular and Clinical Sciences Research Institute, St George's University of London, London SW17 ORE, UK
5 University Institute for Advanced Studies IUSS, 27100, Pavia, Italy
6 IRCCS Mondino Foundation, 27100, Pavia, Italy

Received April 04, 2024. Revised September 10, 2024. Accepted November 27, 2024. Advance access publication November 29, 2024

Correspondence to: Shalom K. Henderson
MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Rd,
Cambridge CB2 7EF, UK
E-mail: Shalom.Henderson@mrc-cbu.cam.ac.uk

Keywords: connected speech; lexico-semantic word properties; picture description word checklist; primary progressive aphasia; Parkinson-plus disorders

Graphical Abstract



Introduction

Speech is an integral part of effective communication and is often disturbed by brain damage such as stroke or neurodegeneration. Breakdown in speech production is important clinically as it can be diagnostic for different types of aphasia. Clinicians use conversations and narratives to detect communication difficulties in people with a speech and/or language impairment. Connected speech elicited by a picture description task, in particular, has been used to distinguish healthy controls from patients with diverse neurodegenerative diseases, as well as between specific subtypes of stroke aphasia and primary progressive aphasia (PPA).[1-3] To aid differential diagnosis and improve understanding about the nature of speech and language changes in PPA, many speech and linguistic measures have been previously investigated (e.g. acoustic/prosodic, lexico-semantic, morpho-syntactic and pragmatic/discourse) and subsequently quantified (e.g.

speech rate, syllable duration, words per minute and psycholinguistic word properties) in connected speech analyses. However, transcription and quantification of speech properties require advanced linguistic expertise and are time-consuming. A simple analytical tool for analysing connected speech would be of great benefit. For example, if a simple target word list can be used (validated by in-depth, systematic analysis of connected speech with high diagnostic differentiation between progressive aphasias), this could be a practical and efficient clinical tool for assessing and diagnosing people with a neurodegenerative language impairment.

An important first step to this objective is to determine the distribution of words produced by each patient group and consider the variety of speech features and psycholinguistic properties. Both qualitative and quantitative differences in connected speech have been reported in PPA. For example, the number of content words is reduced in patients with the semantic variant (svPPA), with over-reliance on highly frequent words; in other words, the content of their speech becomes 'lighter' with overuse of words that are more frequent, less concrete, less imageable and more semantically diverse.[4,5] Even though relatively less is known about the psycholinguistic properties of words produced by the non-fluent (nfvPPA) and logopenic (lvPPA) variants, articulatory and prosodic features, such as syllable duration, speech rate and word length, and grammatical complexity have been reported to differentiate between these two variants.[6-8]

Language impairments are also common in progressive supranuclear palsy (PSP) and corticobasal syndrome (CBS). Both conditions have features that overlap with nfvPPA[9,10] such as dysfluency and syntactic impairments in production and comprehension.[11] Similarities across these three groups have been reported in acoustic and lexical measures of connected speech during a picture description task.[12] Connected speech alterations have been found in PSP patients[13-15] including reduced speech rate, reduced total number of words and sentences, higher number of pronouns and impaired grammatical complexity.[16,17] Only a few studies have investigated connected speech in CBS, with one describing an overall reduction in connectedness (i.e. the number of connected events as a proportion of mentioned events) during a narrative discourse[18] and another reporting reduced speech production rate and lexical-semantic errors during a picture description task.[19]

The differing methods of connected speech analysis in previous investigations pose a challenge in determining which measures, amongst an exhaustive list of word properties and features related to speech/language quantification, are useful for distinguishing between neurodegenerative diseases with a primary or associated language impairment. Here, we sought to address this knowledge gap with the following aims: (i) to determine which speech-related properties differentiate between svPPA, lvPPA, nfvPPA, PSP, CBS and healthy controls during picture description using a principal component analysis (PCA) to understand and simplify the patterns of change in quantifiable speech and psycholinguistic properties of connected speech; (ii) to examine the neural correlates of connected speech in these conditions and (iii) to use a data-driven approach to develop an easy-to-use and practical word checklist.

Materials and methods

Participants

Seventy-four people (24 healthy controls, 9 svPPA, 9 lvPPA, 9 nfvPPA, 10 PSP and 13 CBS) from the Mini Linguistic State Examination (MLSE)[20] study were included in the development data set. Controls were recruited through the National Institute for Health Research 'Join Dementia Research' register and via local advertisement; other participants were recruited from tertiary referral services at Addenbrooke's Hospital, Cambridge ($N = 46$), and Salford Royal Foundation Trust and its associated clinical providers ($N = 4$). Patients from a second site in the MLSE study[20] at St. George's Hospital, London, made an out-of-sample test set with svPPA ($N = 7$), lvPPA ($N = 13$), nfvPPA ($N = 5$), PSP ($N = 2$) and CBS ($N = 6$). Clinical diagnoses of PPA, PSP and CBS were based on current consensus criteria.[21-23] In our development sample, one nfvPPA patient declared a native language of Italian. Two svPPA patients from our out-of-sample site declared a native language of Gujarati and Indian Patois. All three patients who declared a non-English native language were pre-morbidly highly fluent in English.

Connected speech acquisition, transcription, reliability and analysis

Participants completed the MLSE and the Boston Diagnostic Aphasia Examination (BDAE)[24] and were asked to describe both the BDAE 'cookie theft' and MLSE 'beach scene' pictures. The instruction for both pictures was as follows: 'Look carefully at this picture and describe aloud what is happening. Try to use sentences. I will stop you after one minute. Ready?' The examiners politely allowed time for the participants to finish their description after the 1-min mark. Connected speech samples were video recorded and transcribed by a speech-language pathologist (S.K.H.), blinded to the clinical diagnoses, using the f4transcript notation software version 7.0, which has been previously reported to make the manual writing of speech samples from audio or video recordings more efficient. Speech samples were formatted for analysis with the Frequency in Language Analysis Tool (FLAT)[25] which has specific codes for false speech. For example, false starts, grammatical errors, grammatical clause boundaries, prosodic indicators, non-lexical interjections such as filler words and pauses, repetitions, unintelligible segments and neologisms were coded and excluded from the analyses.

To assess transcription reliability, we randomly selected two transcripts from each diagnostic group. We assessed the reliability of two different transcribers (S.K.H. and

K.A.P.) by dividing the number of matching words between the two transcripts by the total words in the transcript used in our analyses (transcribed by the first author, S.K.H.). This method is consistent with previously reported reliability analyses of transcriptions.[26] We found a high per cent agreement between the two transcripts with an average of 92% (range 81–98%) and 100% for the words in the 15-item checklist (below).

Using the transcribed speech samples free of false speech, we calculated the simplest measurements of connected speech (i.e. word counts, ratios and timing) to test whether these can differentiate groups as well as other measures of connected speech that tend to be more time-consuming to score and analyse (e.g. acoustic features). The total number and type counts for words and total time and words per minute were calculated for each participant. Additionally, the number and type counts for word bigrams (i.e. two-word combinations such as 'the mother') and word trigrams (i.e. three-word combinations such as 'sink is overflowing'); type-to-token ratios for words, word bigrams and word trigrams; proportion of function relative to content words; and combination ratio (i.e. a measure of connected language calculated as word trigram count divided by word count)[27] were extracted using an automated script for language quantification with the use of FLAT.[25] These measures of speech fluency for the two-picture description tasks were included in our first PCA.

Next, for the psycholinguistic word properties, each distinct word produced across all participants was extracted for analysis. We then excluded function words (e.g. articles, demonstratives and prepositions), and for each content word, we looked up the ratings from various databases for length, log frequency,[28] semantic diversity,[29] semantic neighbourhood density,[30] concreteness,[31] age of acquisition[32] and orthographic and phonological Levenshtein distance.[33,34] Where ratings for pluralized words were unavailable, word properties for the singular version were extracted. Although ratings for familiarity and imageability were initially obtained, these measures were excluded in the main analysis due to the unavailability of ratings for a high proportion of words. Of the available data, imageability ratings were strongly correlated with concreteness ratings ($R = 0.94$, $P < 0.001$) and familiarity ratings were moderately correlated with log frequency ratings ($R = 0.45$, $P < 0.001$). These word properties were included in our second PCA.

Statistical analysis

We used PCA as a dimensionality reduction method to investigate the distinct speech characteristics underlying connected speech performance. First, we calculated the average counts per participant for the quantifiable measures of speech fluency (e.g. number and type of words, type-to-token ratio and word per minute) for each picture using the transcribed speech samples which were then entered into a varimax-rotated PCA. A Kaiser–Meyer–Olkin test determined the suitability of our data set. We selected three components based on Cattell's criterion. Using principal component (PC) scores per participant, we conducted a one-way analysis of variance (ANOVA) to test for group differences.

Next, to understand the underlying pattern of variations in the lexico-semantic word properties produced by all patients and controls, all unique content words produced by patients and controls in both picture descriptions were compiled into a single 'speech corpus' and the psycholinguistic properties of each word were entered into a varimax-rotated PCA. After selecting three components using Cattell's criteria, PC scores for the words produced by each participant were extracted and then averaged across individual participants. Using these averaged PC scores per participant, we tested the differences between group and task (i.e. 'cookie theft' versus 'beach scene') using a two-way ANOVA. Past studies have found that comparison of mean values can be relatively insensitive for detecting patients' altered word usage, whereas distribution analyses can be more sensitive (e.g. where there are more pronounced changes in one part of the distribution).[4,5] Thus, using data from the word properties PCA, PC scores were split into quartiles (ranging from −4 to −2, greater than −2–0, greater than 0–2 and greater than 2–4). For each participant, we counted the number of times each participant produced words in each range of a PC (e.g. −4 to −2 in PC 1) and each point in the psycholinguistic dimensional space (e.g. −4 to −2 in PC 1 and 2–4 in PC 2). We then generated contour plots that mapped the proportion of words produced by each participant which were then averaged across groups. Using a method previously applied by Hoffman *et al.*,[5] we generated difference plots by subtracting the mean of control data from that of each patient group's data to visualize the differences between control versus patient maps. We explored differences between groups across the variation in word properties in two ways. First, we took the mean value of the proportion of words produced by each patient group and compared them to the control data in each of the dimensional spaces using two-tailed *t*-tests. Secondly, for a more sensitive method, we conducted a distribution analysis by quantifying the number of words produced by controls and patients in each of the PCs' quartiles. A repeated measures ANOVA was performed with quartiles as within-subject and group as between-subject factors.

Post hoc analyses were conducted using Tukey's honestly significant difference test for multiple comparison. All statistical analyses were performed in R statistical software (version 2023.03.0).

Neuroimaging acquisition and voxel-based morphometry analysis

All participants underwent T_1-weighted structural MRI of the brain. Participants from Cambridge were scanned using a 3T Siemens Skyra MRI scanner. Whole-brain T_1-weighted structural images were acquired using the following parameters: iPAT2; 208 contiguous sagittal slices; field of view (FOV) = 282 × 282 mm^2; matrix size 256 ×

256; voxel resolution = 1.1 mm^3; TR/TE/TI = 2000 ms/2.93 ms/850 ms, respectively; and flip angle 8°. Participants from Manchester were scanned using a 3T Philips Achieva MRI scanner. Whole-brain T_1-weighted images were acquired using the following parameters: SENSE = 208 contiguous sagittal slices, FOV = 282 × 282 mm^2, matrix size 256 × 256, voxel resolution = 1.1mm^3, TR/TE/TI = 6600 ms/2.99 ms/850 ms and flip angle 8°.

Whole-brain grey matter changes were indexed using voxel-based morphometry (VBM) analyses of structural T_1-weighted MRI, integrated into Statistical Parametric Mapping software (SPM12: Wellcome Trust Centre for Neuroimaging, https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). A standard pre-processing pipeline was implemented involving (i) brain segmentation into three tissue probability maps (grey matter, white matter and cerebrospinal fluid); (ii) normalization (using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra, DARTEL);[35] (iii) study-specific template creation using grey matter tissue probability maps; (iv) spatial transformation to Montreal Neurological Institute space using transformation parameters from the corresponding DARTEL template; and (v) image modulation and smoothing using 8 mm full-width-half-maximum Gaussian kernel to increase signal-to-noise ratio. Segmented, normalized, modulated and smoothed grey matter images were used for VBM analyses.

We examined the associations between whole-brain grey matter intensity and PCA-generated PC scores, which were averaged across two-picture description tasks, using *t*-contrasts. Age and total intracranial volume were included as nuisance covariates. Clusters were extracted using a threshold of $P < 0.001$ uncorrected for multiple comparisons with a cluster threshold of 100 voxels. We chose 100 voxels as our cluster threshold as we were interested in smaller subcortical regions that have been reported to be associated with speech production and are often atrophic in the disease groups.

Word checklist analysis

To determine target words that could best differentiate between groups, we used least absolute shrinkage and selection operator (LASSO) logistic regression.[36] Given the large number of predictors (i.e. 500+ unique words used by the whole group), relatively small sample size per group and multicollinearity of the words (e.g. the likelihood that a participant would say 'overflowing' and 'sink'), the LASSO method is highly appropriate for automated feature selection and shrinkage. While multiple correlated words are entered into the model, only the most important predictor variables (i.e. the least number of words that best differentiate between groups) will be selected. As a first step, we pooled together all of the different words that patients and controls produced which resulted in over 500+ tokens per picture. We then streamlined this collection by carrying out LASSO regressions for each picture including all unique words produced per picture as predictors for the following comparisons: (i) controls versus each patient group and (ii) each patient group against one another. Whether or not a participant produced a word such as 'overflowing' was coded as 1 for produced and 0 for not produced. We accounted for differences in dialect (e.g. score of 1 if the participant said boy, chap, lad or bloke) and morpho-syntax such as verb tense (e.g. stealing/stolen) and singular/plural forms (e.g. plate/plates).

Next, the words that had been selected in each pairwise comparison (by logistic LASSO regression) were compiled, resulting in a pool of 33 words for the BDAE 'cookie theft' and 46 words for the MLSE 'beach scene' pictures. We re-ran the LASSO regressions for each pairwise comparison using these truncated lists, and the resulting words were further rank ordered by (i) the number of times they appear in the pairwise comparisons, (ii) their beta coefficients and (iii) the magnitude of difference in the overall proportion by group (e.g. magnitude would be 1 if all of the controls produced the word 'overflow' but none of the svPPAs did). The top 15 words resulting from this rank ordering were entered into a series of 4-fold cross-validated LASSO logistic regressions with each predicting the diagnostic distinction of interest (e.g. controls versus patients). The scoresheets using the 15 words are shown in Supplementary Appendix 1.

To evaluate the robustness of the model in predicting group classification with the word checklists, we conducted out-of-sample predictive validity testing with connected speech data from St. George's Hospital. There were no differences in demographics between patients from the two test sites except for PSP patients from St. George's having lower scores on the Addenbrooke's Cognitive Examination Revised (ACE-R) compared to those from Cambridge ($P = 0.02$). We tested the 15-word checklist with the St. George's data assigning a score of 1 if the participant produced the target word and a 0 if the word was omitted. Morpho-syntactic variations were scored as correct if the root matched the target word (e.g. overflowing for overflow and digging for dig). As an index of accuracy for our binomial models (i.e. pairwise comparisons), we report classification performance on the test data using function confusion.glmnet from the glmnet package in R for the following comparisons: controls versus all patients, patients belonging to the 'motor' group (i.e. nfvPPA, PSP and CBS) versus 'lexico-semantic' group (i.e. svPPA, lvPPA) and each patient group against one another. Of note, PSP and CBS patients were grouped into one due to small sample size (i.e. two PSP) in our out-of-sample test set.

To test the hypothesis that supplementing the checklist with cognitive scores might improve the differentiation between groups, we ran another LASSO logistic regression with the 15 words (coded the same way as noted above), as well as subtest scores from the ACE-R and MLSE. We estimated the LASSO model using a within-sample 4-fold cross-validation with the Cambridge training set and tested the generalizability of our model with the St. George's data as out-of-sample test.

Results

Demographics

Demographic and clinical features are shown in Table 1. There were no significant differences in all groups for age, gender and handedness, as well as symptom duration for patients. There were significant differences between groups in education; *post hoc* tests confirmed that controls left education later than patients with nfvPPA, CBS and PSP ($P < 0.05$). Significant group differences emerged on total MLSE and ACE-R scores. Controls performed better on the MLSE when compared with patients with svPPA, lvPPA, nfvPPA and CBS ($P < 0.001$), PSP performed better than lvPPA ($P = 0.001$) and nfvPPA ($P = 0.007$) and CBS performed better than lvPPA ($P = 0.03$). On the ACE-R, controls performed better than all patient groups ($P < 0.05$); nfvPPA, PSP and CBS performed better than lvPPA ($P < 0.05$); and PSP performed better than svPPA ($P = 0.001$). As shown in Table 1, in our development sample, all participants were white. Two svPPA patients from our out-of-sample site were non-white.

Quantification of speech fluency

Average counts per participant for the quantifiable properties of words and word combinations per picture were entered into a PCA with varimax rotation. Three PCs were identified using Cattell's criteria which explained 86.5% of the variance (Kaiser–Meyer–Olkin = 0.70). The loadings of each measure are shown in Supplementary Table 1.

Type and token counts for words, word bigrams and word trigrams; word per minute; type-to-token ratio of words; and combination ratio loaded most heavily on PC 1, and thus, we labelled this PC as 'speech quanta'. Type-to-token ratio of words, word bigrams and word trigrams loaded most heavily on PC 2 which we labelled as 'lexical richness'. Word per minute, an index of speech fluency and combination ratio, the degree to which an individual produced longer, more complex combinations as opposed to single word fragments, loaded heavily on PC 3, and we adopted the working label of 'speech complexity'.

Group performance patterns on all three PCs are visually summarized in Fig. 1A. For PC 1, the results from a one-way ANOVA revealed group differences [$F(1142) = 71.19$, $P < 0.001$], driven by controls and svPPA patients having higher scores than those with nfvPPA ($P < 0.001$), PSP ($P < 0.01$) and CBS ($P < 0.05$). Additionally, controls had higher scores than patients with lvPPA ($P = 0.01$), who in turn had higher scores than those with nfvPPA ($P < 0.001$). A one-way ANOVA did not reveal group differences for PC 2 [$F(1142) = 1.26$, $P = 0.26$]. For PC 3, the results from a one-way ANOVA revealed group differences [$F(1142) = 12.77$, $P < 0.001$], driven by controls having higher scores than those with nfvPPA ($P < 0.001$), PSP ($P < 0.001$) and CBS ($P = 0.002$).

Correlations between the speech fluency PC scores and total and subdomain scores of the MLSE can be found in Supplementary Table 2.

Quantification of word properties

Ratings of psycholinguistic features for all words produced by controls and patients were entered into a PCA with varimax rotation. Three PCs were identified using Cattell's criteria, each representing a group of covarying psycholinguistic features. These three components explained 85.5% of the variance (Kaiser–Meyer–Olkin = 0.75). The loadings of each measure are shown in Supplementary Table 3. Length and phonological and orthographic Levenshtein distance loaded most heavily on PC 1, and we adopted the working label of 'length'. Concreteness, log frequency, semantic neighbourhood density and semantic diversity loaded heavily on PC 2 which we labelled as 'semantic richness'. Age of acquisition loaded most heavily on PC 3 which we labelled as 'acquisition age'.

The three scores, obtained from the psycholinguistic PCA results, per participant along with the elicitation task were into a two-way ANOVA which revealed significant group differences in PC 1 [$F(5134) = 4.29$, $P < 0.001$], driven by svPPA and lvPPA patients producing words that were shorter, phonologically and orthographically less complex than controls ($P < 0.05$) (see Fig. 1B).

Table 1 Demographics and clinical features of the study cohort

	Control	svPPA	lvPPA	nfvPPA	PSP	CBS	*P**
N	24	9	9	9	10	13	-
Age (SD)	65.8 (5.2)	67.2 (4.3)	68.9 (8.1)	70.1 (6.4)	68.4 (5.9)	70.2 (4.4)	ns
Gender M:F	11:13	5:4	6:3	4:5	5:5	7:6	ns
Handedness R:L	21:3	9:0	9:0	8:1	9:1	12:1	ns
Ethnicity white:other	24:0	9:0	9:0	9:0	9:0	9:0	-
First language English:other	24:0	9:0	9:0	8:1	9:0	9:0	-
Age left education (SD)	20.6 (3.3)	19.3 (2.6)	20.6 (4.1)	16.6 (1.7)	16.7 (1.7)	17.4 (3.0)	<0.001
Symptom duration in years (SD)	NA	6.5 (2.5)	3.0 (2.7)	3.2 (2.9)	4.1 (2.5)	5.2 (4.0)	ns
Total MLSE (SD)	98.3 (2.2)	78.1 (4.7)	68.1 (15.3)	70.9 (15.5)	87.9 (8.0)	81.8 (14.3)	<0.001
ACE-R (SD)	96.0 (3.4)	53.9 (8.2)	46.7 (25.1)	69.7 (15.1)	80.5 (13.4)	74.0 (17.6)	<0.001

Note: Mean and standard deviations are displayed. For MLSE and ACE-R, values indicate scores out of 100. **P*-value for *F*-test of group difference by ANOVA. ACE-R, Addenbrooke's Cognitive Examination Revised; CBS, corticobasal syndrome; lvPPA, logopenic variant primary progressive aphasia; MLSE, Mini Linguistic State Examination; nfvPPA, non-fluent variant primary progressive aphasia; ns, not significant, $P > 0.1$; PSP, progressive supranuclear palsy; SD, standard deviation; svPPA, semantic variant primary progressive aphasia.

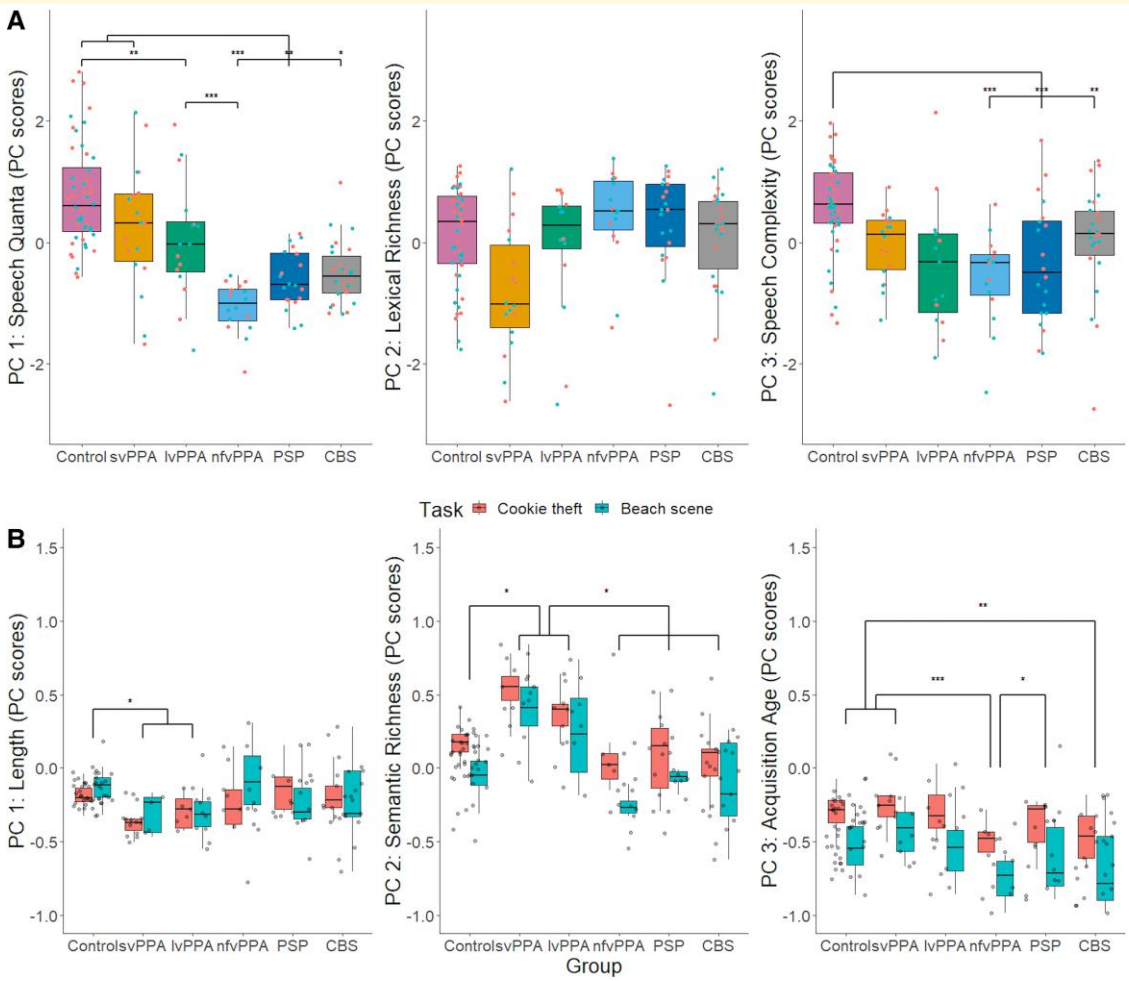


Figure 1 PCA scores across diagnostic groups. (**A**) Scores of quantitative measures of speech fluency. For PC 1 ('speech quanta'), the results from a one-way ANOVA revealed significant group differences [$F(1142) = 71.19$, $P < 0.001$], driven by controls ($N = 24$) and patients with svPPA ($N = 9$) having higher scores than those with nfvPPA ($N = 9$), PSP ($N = 10$) and CBS ($N = 13$), controls having higher scores than those with lvPPA ($N = 9$) and patients with lvPPA having higher scores than those with nfvPPA. PC 2 ('lexical richness') resulted in no group differences [$F(1142) = 1.26$, $P = 0.26$], and for PC 3 ('speech complexity'), significant group differences were found [$F(1142) = 12.77$, $P < 0.001$], driven by controls having higher scores than patients with nfvPPA ($P < 0.001$), PSP ($P < 0.001$) and CBS ($P = 0.002$). (**B**) Scores of quantitative measures of word properties across groups. For PC 1 ('length'), the results from a two-way ANOVA revealed significant group differences [$F(5134) = 4.29$, $P < 0.001$], driven by svPPA and lvPPA patients producing words that were shorter, phonologically and orthographically less complex than controls ($P < 0.05$). For PC 2 ('semantic richness'), significant differences were found for group [$F(5134) = 16.62$, $P < 0.001$] and task [$F(1134) = 22.05$, $P < 0.001$]. Patients with svPPA and lvPPA produced more words that were characterized as more frequent and semantically diverse than those with nfvPPA, PSP, CBS and controls ($P < 0.01$). For PC 3 ('acquisition age'), significant differences were found for group [$F(5134) = 7.09$, $P < 0.001$] and task [$F(1134) = 50.24$, $P < 0.001$]. *Post hoc* analyses revealed that (i) nfvPPA patients produced words that were characterized as significantly earlier acquired than those with svPPA ($P < 0.001$), PSP ($P = 0.05$) and controls ($P < 0.001$) and (ii) CBS patients used words that were significantly earlier acquired than those with svPPA ($P = 0.002$) and controls ($P = 0.01$). Results from *post hoc* analyses using Tukey's honestly significant difference test for multiple comparisons are shown as asterisks indicating level of significance: *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$. CBS, corticobasal syndrome; lvPPA, logopenic variant of primary progressive aphasia; nfvPPA, non-fluent variant of primary progressive aphasia; PC, principal component; PSP, progressive supranuclear palsy; svPPA, semantic variant of primary progressive aphasia.

For PC 2, significant differences were found for group [$F(5134) = 16.62$, $P < 0.001$] and task [$F(1134) = 22.05$, $P < 0.001$]. The task effect was driven by more frequent and semantically diverse words produced for the 'cookie theft' than the 'beach scene' picture. *Post hoc* analyses revealed that svPPA and lvPPA patients produced more words that were characterized as more frequent and semantically diverse than those with nfvPPA, PSP, CBS and controls ($P < 0.01$).

Significant differences were found for group [$F(5134) = 7.09$, $P < 0.001$] and task [$F(1134) = 50.24$, $P < 0.001$] for PC 3. The words used to describe the 'cookie theft' were found to be later acquired. *Post hoc* analyses revealed that nfvPPA patients produced words that were characterized as significantly earlier acquired than those with svPPA ($P < 0.001$), PSP ($P = 0.05$) and controls ($P < 0.001$). Similarly, CBS patients used words that were significantly

earlier acquired than those with svPPA ($P = 0.002$) and controls ($P = 0.01$).

Correlations between the word properties PC scores and total and subdomain scores of the MLSE can be found in Supplementary Table 2.

Differences in multivariate word properties

Moving beyond the simplistic mean statistic, we looked at the bivariate distributions of words across the psycholinguistic space and how these might shift in each patient group (e.g. patients produce fewer words in one part of the space and might substitute more words in another part of the space). Figure 2 shows the contour plot for controls (left), depicting the averaged proportion of words produced within the PC space, and the difference plots where the mean of the control data for the three PCs from the 'word properties' PCA were subtracted from that of the patient data.

Relative to controls, svPPA and lvPPA patients produced a greater proportion of words in the higher semantic richness (i.e. more semantically diverse and frequent) and lower length (i.e. shorter, less phonologically and orthographically complex) space. In contrast, nfvPPA, PSP and CBS patients produced a greater proportion of words with lower semantic richness and acquisition age (i.e. earlier acquired) space.

Distribution analysis of word properties PCA

Another way to go beyond the simplistic mean statistic is to undertake a formal distribution analysis for each PC. This has been shown in previous work to be much more sensitive to changes in the content words produced by patients.[37,38] As shown in Fig. 3, PC scores for PC 1 to PC 3 from the word properties PCA were divided into quartiles and the number of words produced in each quartile was computed for each participant followed by a group mean.

For PC 1, a six groups × four quartiles repeated measures ANOVA showed a significant effect of group only for both 'cookie theft' [$F(5283) = 37.16$, $P < 0.001$] and 'beach scene' [$F(5272) = 39.18$, $P < 0.001$]. For PC 2, a six groups × four quartiles repeated measures ANOVA showed significant effects of group [$F(5280) = 33.68$, $P < 0.001$], quartile [$F(1280) = 4.67$, $P = 0.03$] and group-by-quartile interaction [$F(5280) = 4.36$, $P < 0.001$] for 'cookie theft'. For 'beach scene', a six groups × four quartiles repeated measures ANOVA showed significant effects of group [$F(5270) = 28.94$, $P < 0.001$], quartile [$F(1270) = 5.53$, $P = 0.02$] and group-by-quartile interaction [$F(5270) = 8.29$, $P < 0.001$]. For PC 3, a six groups × four quartiles repeated measures ANOVA showed significant effects of group [$F(5283) = 36.15$, $P < 0.001$], quartile [$F(1283) = 17.17$, $P < 0.001$] and

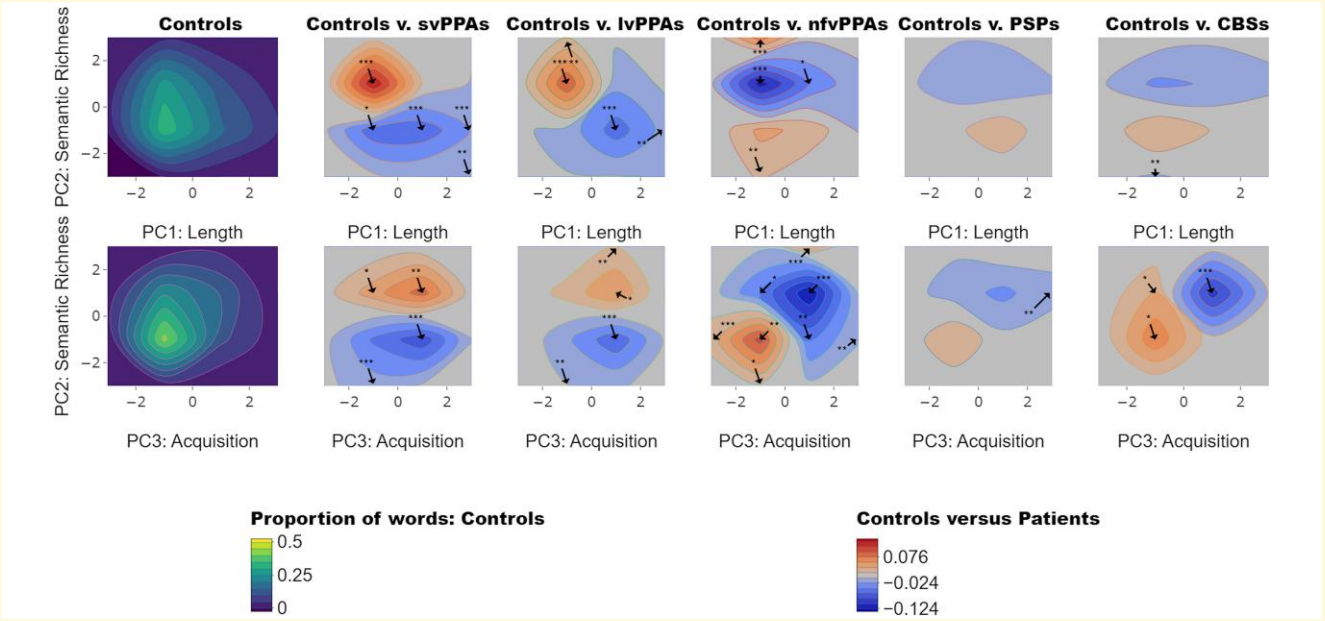


Figure 2 Contour distributions and difference plots. The top and bottom left plots show the contour distributions across PC 1 (length), PC 2 (semantic richness) and PC 3 (acquisition age) produced by healthy controls. Difference plots comparing patients with healthy controls are shown to the right of the contour plots of healthy controls. In the control plots, yellow tones show where the greatest proportions of words were found within the PC space. For controls versus patients, the red and blue tones represent PC spaces where patients produced more words than controls and where controls produced more than patients, respectively. Taking the mean value of the proportion of words produced by each patient group (svPPA $N = 9$, lvPPA $N = 9$, nfvPPA $N = 9$, PSP $N = 10$ and CBS $N = 13$), we compared them to the control data ($N = 24$) in each of the dimensional spaces using two-tailed *t*-tests. The arrows indicate where in the maps there were significant differences between controls and patients (*P*-values are shown as asterisks indicating level of significance: *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). CBS, corticobasal syndrome; lvPPA, logopenic variant of primary progressive aphasia; nfvPPA, non-fluent variant of primary progressive aphasia; PSP, progressive supranuclear palsy; svPPA, semantic variant of primary progressive aphasia.

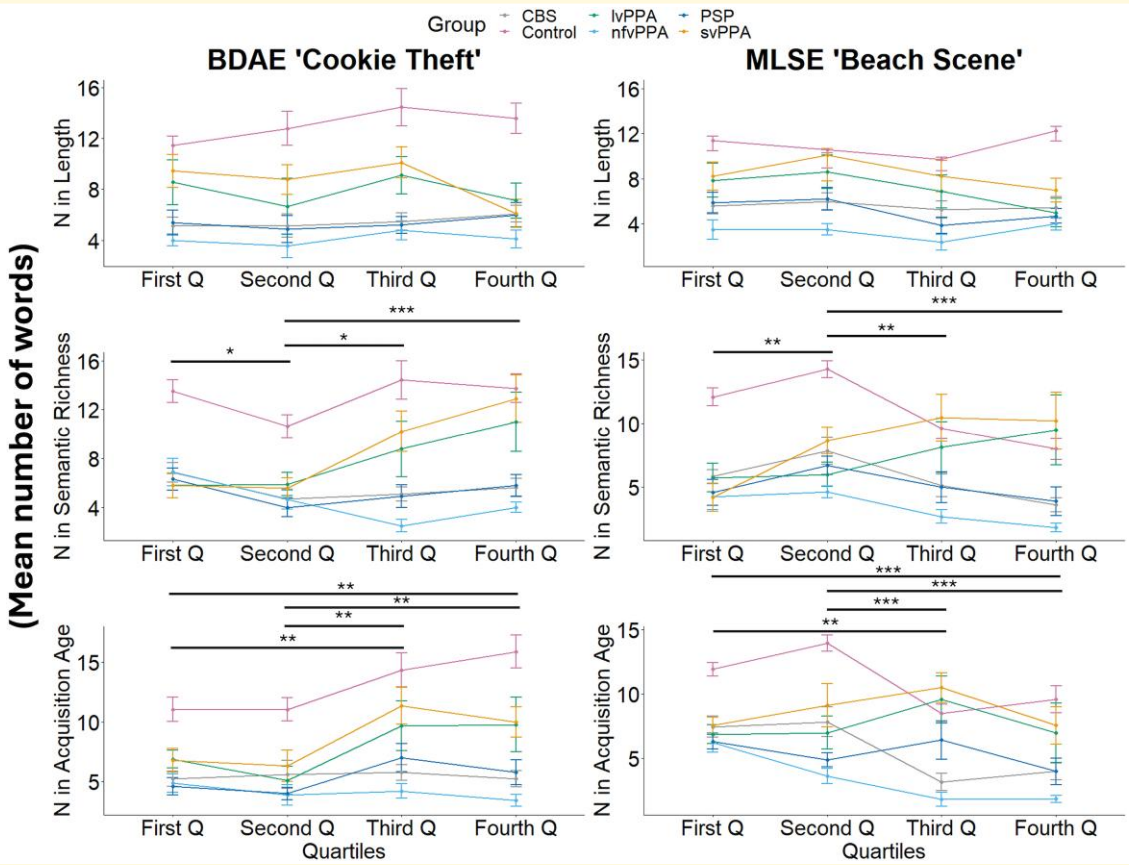


Figure 3 Distribution plots. Each plot shows the mean number of words produced in each quartile (Q) by patient groups (svPPA $N = 9$, lvPPA $N = 9$, nfvPPA $N = 9$, PSP $N = 10$ and CBS $N = 13$) for PC 1 'length', PC 2 'semantic richness' and PC 3 'acquisition age'. For PC 1, a six groups × four quartiles repeated measures ANOVA showed a significant effect of group for both 'cookie theft' [$F(5283) = 37.16$, $P < 0.001$] and 'beach scene' [$F(5272) = 39.18$, $P < 0.001$]. For PC 2, significant effects were found for group [$F(5280) = 33.68$, $P < 0.001$], quartile [$F(1280) = 4.67$, $P = 0.03$] and group-by-quartile interaction [$F(5280) = 4.36$, $P < 0.001$] for 'cookie theft'. For 'beach scene', significant effects were found for group [$F(5270) = 28.94$, $P < 0.001$], quartile [$F(1270) = 5.53$, $P = 0.02$] and group-by-quartile interaction [$F(5270) = 8.29$, $P < 0.001$]. For PC 3, significant effects were found for group [$F(5283) = 36.15$, $P < 0.001$], quartile [$F(1283) = 17.17$, $P < 0.001$] and group-by-quartile interaction [$F(5283) = 2.47$, $P = 0.03$] for 'cookie theft'. For 'beach scene', significant effects were found for group [$F(5265) = 31.04$, $P < 0.001$], quartile [$F(1265) = 21.67$, $P < 0.001$] and group-by-quartile interaction [$F(5265) = 2.47$, $P = 0.03$]. The effect of quartile from *post hoc* analyses using Tukey's honestly significant difference test for multiple comparisons is shown as asterisks indicating level of significance: *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$. BDAE, Boston Diagnostic Aphasia Examination; CBS, corticobasal syndrome; lvPPA, logopenic variant of primary progressive aphasia; MLSE, Mini Linguistic State Examination; nfvPPA, non-fluent variant of primary progressive aphasia; PSP, progressive supranuclear palsy; svPPA, semantic variant of primary progressive aphasia.

group-by-quartile interaction [$F(5283) = 2.47$, $P = 0.03$] for 'cookie theft'. For 'beach scene', a six groups × four quartiles repeated measures ANOVA showed significant effects of group [$F(5265) = 31.04$, $P < 0.001$], quartile [$F(1265) = 21.67$, $P < 0.001$] and group-by-quartile interaction [$F(5265) = 2.47$, $P = 0.03$]. Our results are summarized in Supplementary Table 4.

Neural correlates of connected speech properties

Associations between grey matter intensity and PC scores from both quantitative measures of speech fluency and word properties are shown in Fig. 4 and Supplementary Table 5. In the entire group (i.e. patients and controls), PC 1 ('speech quanta') scores correlated with grey matter intensities of the bilateral middle and superior frontal gyri, right inferior frontal gyrus (IFG), insula, putamen and caudate. PC 3 ('speech complexity') scores correlated with grey matter intensities of the left insula; inferior, middle and superior frontal gyri, extending medially; superior temporal gyrus (STG); and parts of the limbic system. No significant correlations were found for PC2 ('lexical richness') scores.

For the word properties PCA, PC 1 ('length') scores correlated with grey matter intensities of the left insula, middle and superior temporal gyri, bilateral parahippocampal and fusiform gyri, right inferior and middle temporal gyri and limbic structures. PC 3 ('acquisition age') scores correlated

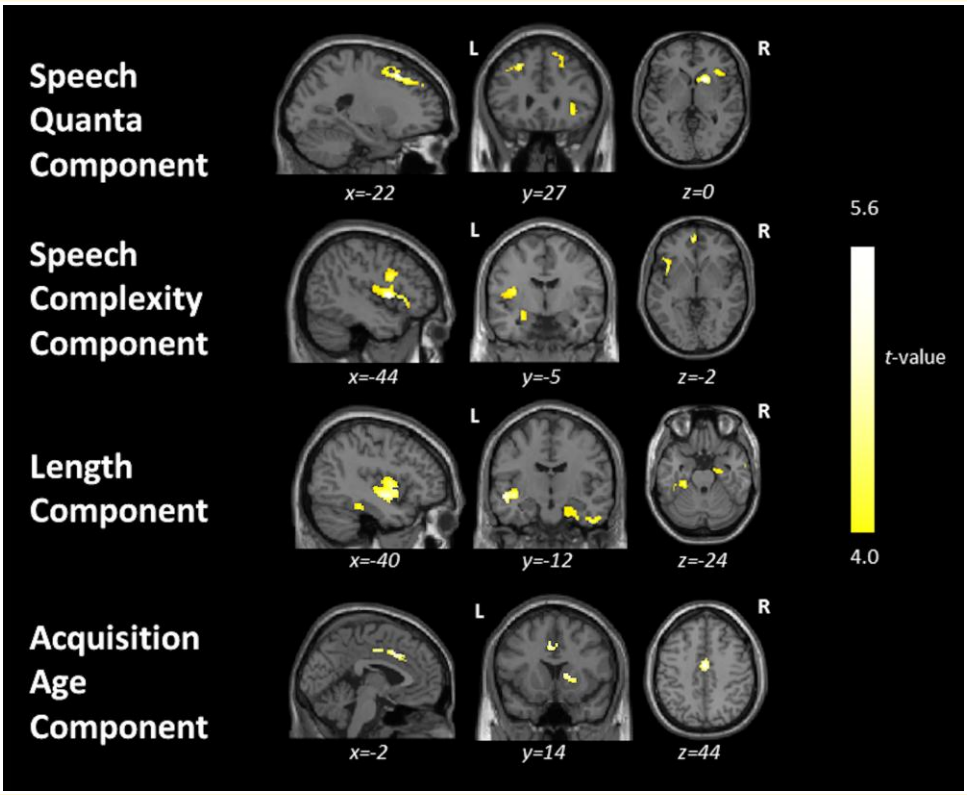


Figure 4 Results from the whole-brain VBM correlation analyses. This figure shows regions of grey matter intensity that uniquely correlate with PC scores in the whole group including controls ($N = 24$) and patients (svPPA $N = 9$, lvPPA $N = 9$, nfvPPA $N = 9$, PSP $N = 10$ and CBS $N = 13$) using *t*-contrasts. Clusters were extracted using a threshold of $P < 0.001$ uncorrected for multiple comparisons with a cluster threshold of 100 voxels with age and total intracranial volume included as nuisance covariates. CBS, corticobasal syndrome; lvPPA, logopenic variant of primary progressive aphasia; nfvPPA, non-fluent variant of primary progressive aphasia; PSP, progressive supranuclear palsy; svPPA, semantic variant of primary progressive aphasia.

with grey matter intensities of the bilateral cingulate gyri and right caudate and putamen. No significant correlations were found for PC 2 ('semantic richness') scores.

When excluding healthy controls, PC 1 ('length') scores correlated significantly with a single cluster including the left insula and middle and superior temporal gyri (see Supplementary Table 6). No significant correlations were found for the other PC scores. Supplementary Table 7 shows the results when using a cluster-forming height threshold of $P < 0.005$ paired with a cluster extent threshold of $P < 0.05$ FWE-corrected.

Word checklist

Using the word checklist for each picture (see Supplementary Appendix 1), the LASSO logistic regression selected a group of words that together predicted group membership (see Supplementary Table 8). The checklist scoresheets for additional diagnostic differentiations (e.g. svPPA versus nfvPPA, PSP and CBS) and a representative example of an anonymised patient using the BDAE 'cookie theft' 15-word checklist scoresheet can be found in Supplementary Appendix 2 and Supplementary Table 9, respectively. Of note, the LASSO regression for svPPA versus lvPPA and nfvPPA versus PSP resulted in zero words for both pictures; in other words, none of these words could differentiate between these groups. These results motivated our hierarchical classification as shown in Fig. 5, where the 'motor' group included patients with nfvPPA, PSP and CBS and the 'lexico-semantic' group included those with svPPA and lvPPA. The within-sample *k*-fold validation accuracies for 'cookie theft' were as follows: 96% for patients versus controls and 92% for 'motor' versus 'lexico-semantic' groups. Out-of-sample test accuracy with the St. George's data ($N = 34$) resulted in 91% for patients versus controls and 74% for 'motor' versus 'lexico-semantic' groups.

For 'beach scene', the within-sample *k*-fold validation accuracies were as follows: 94% for patients versus controls and 88% for 'motor' versus 'lexico-semantic' groups. Out-of-sample test accuracy resulted in 97% for patients versus controls and 59% for 'motor' versus 'lexico-semantic' groups. Of note, the LASSO regression for nfvPPA versus PSP and CBS combined also resulted in zero words for both pictures.

Since we were not able to differentiate individual patient groups using the checklist alone, we tested the hypothesis

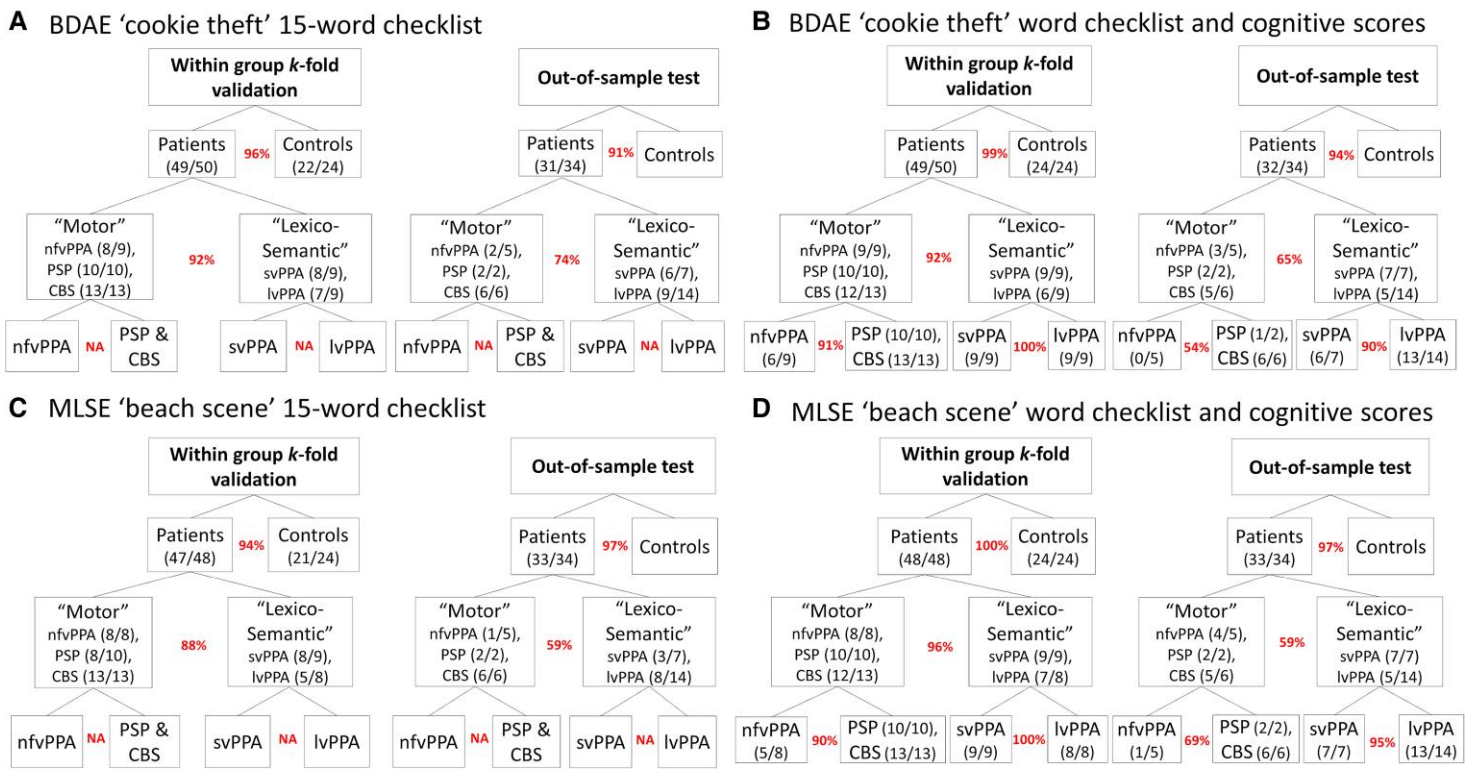


Figure 5 Within-sample *k*-fold and out-of-sample validations. This figure summarizes the validations for (**A**) BDAE 'cookie theft' 15-word checklist, (**B**) BDAE 'cookie theft' 15-word checklist with cognitive measures of ACE-R and MLSE, (**C**) MLSE 'beach scene' 15-word checklist and (**D**) MLSE 'beach scene' 15-word checklist with cognitive measures of ACE-R and MLSE. The curved brackets under or next to the group names (e.g. Patients, 'Motor' and 'Lexico-Semantic') illustrate the number of participants who were classified correctly out of the total *N*. The percentages in red indicate the within-sample 4-fold and out-of-sample validation accuracies. ACE-R, Addenbrooke's Cognitive Examination Revised; BDAE, Boston Diagnostic Aphasia Examination; CBS, corticobasal syndrome; lvPPA, logopenic variant of primary progressive aphasia; MLSE, Mini Linguistic State Examination; nfvPPA, non-fluent variant of primary progressive aphasia; PSP, progressive supranuclear palsy; svPPA, semantic variant of primary progressive aphasia.

that supplementing with cognitive measures might improve the differentiation between these groups. To this end, we supplemented the LASSO models with ACE-R and MLSE sub-scores along with the target words and found improved differentiation for within-sample validation for both nfvPPA versus PSP and CBS (91% for 'cookie theft' and 90% for 'beach scene') and svPPA versus lvPPA groups (100% for both 'cookie theft' and 'beach scene'). Moreover, results from the out-of-sample predictive validity testing showed that the checklists and LASSO models were generalizable more for svPPA versus lvPPA (90% for 'cookie theft' and 95% for 'beach scene') when compared with nfvPPA versus PSP and CBS (54% for 'cookie theft' and 69% for 'beach scene').

In Fig. 5, the curved brackets under or next the group names (e.g. Patients, 'Motor' and 'Lexico-Semantic') illustrate the number of participants who were classified correctly. Supplementary Fig. 1 shows each participant's scores on the 15-word checklists and ACE-R and highlights the participants who were misclassified. The sensitivity and specificity of the word checklists are shown in Supplementary Fig. 2.

Discussion

Clinical impressions from listening to patients' speech are often used to guide diagnosis, but there are two main challenges that this study addresses. First, it is not clear what aspects of the speech should be the target of the assessment. Second, although samples of speech are easy to collect, detailed analyses of connected speech are time-consuming and require specialist expertise. In the present study, we undertook detailed transcription and analyses of connected speech elicited by two-picture description tasks and established which speech features and/or psycholinguistic properties might show the greatest differentiation across groups. We then identified the atrophy correlates of speech-related features. Finally, using data-driven methods, we established a clinically efficient and effective vocabulary checklist method to aid differential diagnosis between the subtypes of PPA, PSP and CBS.

We found significant differences in both speech features and psycholinguistic properties of words between patients and controls. These features also differentiated svPPA and lvPPA versus the remaining groups which are most typically associated with a tauopathy and/or motor disorders (nfvPPA, CBS and PSP). The total language output was significantly reduced in patients with nfvPPA, PSP and CBS relative to those with svPPA and controls. Inspection of the proportion of words produced across the lexico-semantic space revealed that patients with svPPA and lvPPA used a greater proportion of words with high semantic richness (i.e. more frequent and semantically diverse) and lower length (i.e. shorter, less phonologically and orthographically complex) such as 'do', 'out' and 'get' relative to controls. In contrast, patients with nfvPPA, PSP and CBS showed the opposite pattern with a greater proportion of words in the lower semantic richness and acquisition age (i.e. earlier acquired) space such as 'dog', 'boy' and 'cookie'.

We demonstrated that a straightforward word checklist can provide a 'user-friendly' tool, quantifiable in a simple way, with high sensitivity in differentiating healthy controls from patients with a progressive aphasia. The 15-word checklist showed excellent accuracy for within-sample *k*-fold validation, for differentiating patient groups from controls. Even on an out-of-sample validation data set, the 15-word checklist was excellent at differentiating patients from controls (out-of-sample test accuracy of 91% and 97% for 'cookie theft' and 'beach scene') and moderately good at differentiating primary 'lexico-semantic' (svPPA and lvPPA) from 'motor' (nfvPPA, PSP and CBS) groups (accuracy of 74% and 59% for 'cookie theft' and 'beach scene'). The 15 words did not accurately differentiate patients with svPPA from lvPPA, or nfvPPA from PSP and CBS. This is perhaps unsurprising given the patients' similar patterns of word usage, total language output and psycholinguistic properties of the words elicited. Supplementing the 15-word checklist with cognitive measures of ACE-R and MLSE subtest scores increased diagnostic accuracy for nfvPPA versus PSP and CBS for within-sample validation (91% for 'cookie theft' and 90% for 'beach scene'), as well as svPPA versus lvPPA for both within-sample (100% for both 'cookie theft' and 'beach scene') and out-of-sample validation (90% for 'cookie theft' and 95% for 'beach scene'). With regard to differentiating patients from controls, the best ACE-R subtest was verbal fluency which replicates a recent study that found this simple clinical assessment is excellent at differentiating patients from controls but has limited use for differential diagnosis between patient subgroups.[39] We propose that the quick and simple 15-word checklist is a suitable screening test to identify people with progressive aphasia, although further specialist assessment is needed for accurate diagnostic subtyping. In the following sections, we interpret these findings, consider their clinical implications and note directions for future research.

Reduced language output from nfvPPA, PSP and CBS

Patients with nfvPPA, PSP and CBS were distinguishable from those with svPPA, lvPPA and controls, based on reduced language output and connected speech fluency (as measured by the 'speech quanta' and 'speech complexity' PCs). In particular, combination ratio has been previously proposed as a measure of connected language output because it represents the degree to which an individual produces longer, more complex combinations of words over the total word count.[27] Many studies have suggested that measures such as reduced language output, slowed articulation rate, speech sound errors and proportion of function to content words can differentiate patients with nfvPPA from the other variants of PPA in connected speech and other language tasks.[2,40-44] Interestingly, even without measures of acoustics/prosody such as speech pauses, articulation rate

and syllable duration (that are technically difficult to code and quantify), we were able to differentiate between nfvPPA, PSP and CBS versus svPPA, lvPPA and controls using a simple quantification of connected speech (e.g. type/token count).

Despite a sparse literature on connected speech in PSP and CBS, reduced language output and speech rate have been reported in both groups.[12,17,19] In the present study, PSP and CBS patients were comparable to nfvPPA patients in that all groups produced fewer words with reduced speech complexity. Our results support previous findings[19,45] that a general reduction in language output may be a characteristic pattern of PSP and CBS patients, like those with nfvPPA. Moreover, overall performance on various cognitive and language assessments has also been reported to be similar for PSP, CBS and nfvPPA patients.[11,20,46]

Lexico-semantic features

SvPPA and lvPPA patients produced a greater proportion of words that are more frequent and semantically diverse, as well as shorter and less phonologically complex. This finding is consistent with previous reports and highlights two important points.[4,7] First, the secondary changes in other psycholinguistic properties such as imageability and length may be related to the under-sampling of the low frequency words used by controls; in other words, svPPA patients generated more 'lighter' words that tend to be less imageable and more semantically diverse (e.g. 'something'). In addition to under-sampling the low frequency space, svPPA patients have also been found to over-sample the higher frequency space by substituting alternatives to the low frequency target items or picture elements they are unable to name.[4] For example, in the present study, svPPA patients tended to replace low frequency words typically produced by controls (e.g. 'the *sink* is *overflowing*') with higher frequency words that are less imageable and shorter (e.g. '*it*'s *coming* out').[47] Additionally, prior studies have consistently reported that patients with svPPA/semantic dementia replace content words with high frequency, high semantic diversity and low imageability words not only during picture description, but also in other aspects of language output such as naming and verbal fluency.[4,5,8,39,48,49] Frequency and age of acquisition effects in svPPA have also been found beyond tests requiring language output such as lexical decision.[50] Less is known about the psycholinguistic properties of words used by patients with lvPPA. Our findings accord with those of Cho *et al.*[51] who reported that lvPPA patients produced shorter and more frequent content words when describing the 'cookie theft' picture. Furthermore, our formal distribution analysis with the difference plots (Fig. 2) and quantification of words produced in each quartile (Fig. 3) revealed contrastive patterns across the patient groups with (i) svPPA and lvPPA producing shorter words with high frequency and semantic diversity and (ii) nfvPPA, PSP and CBS producing later acquired, lower frequency and less semantically diverse words.

Grey matter correlates of connected speech features

High scores on the 'speech quanta' PC correlated with greater grey matter intensities of bilateral middle and superior frontal gyri and right IFG extending medially and subcortically to include the insula. Cho *et al.*[40] found increased speech errors and production of partial words in nfvPPA to be associated with cortical thinning in the left middle frontal gyrus. Ash *et al.*[2] found speech sound deficits and reduced speech rate in nfvPPA to be related to atrophy in the insula, a region thought to be important for speech articulation,[52,53] and right premotor and supplementary motor regions. Prior studies have also suggested the role of the superior and middle frontal gyri in the grammatical processing of language production and comprehension.[54,55] These findings highlight the potential role of the bilateral frontal region in measures of speech production and rate.

High scores on the 'speech complexity' PC correlated with grey matter intensities of the left insula, IFG, STG and limbic structures. The largest cluster was found for the left insula and IFG, extending into the temporal lobe. Beyond overt speech production, the IFG and insula are reported to be critical in the acoustic measures of speech production such as pause segment duration in motor speech disorders including nfvPPA, ALS and post-stroke aphasia.[52,56,57] Our findings are in line with previously reported associations between superior temporal regions and greater morpho-syntactic demands,[58] grammaticality,[2] complex sentence production,[59] lexical phonology[60] and verbal generation in controls and diverse patient groups.[45] The STG has also been reported to be implicated in the prefrontal-temporal feedback loop and associated with self-monitoring of speech output.[61]

High scores on the 'length' PC correlated with greater grey matter intensities of the bilateral temporal lobe, including medial temporal regions, insula and right limbic lobe. Notably, when excluding controls, the only cluster that correlated significantly included the left insula and middle and superior temporal gyri (see Supplementary Table 6). During an overt picture naming task, Wilson *et al.*[62] found word length to be positively correlated with signal intensity in the left STG in healthy controls. In addition, Hodgson *et al.*[63] found the middle and superior temporal regions to be not only implicated in phonology but also general semantics and semantic control. The ability to generate longer, phonologically more complex words and word combinations may rely on processing speech sounds, as well as accessing conceptual knowledge and controlled retrieval of meaningful semantic information.

Word checklist for picture description

Validated tools to analyse connected speech samples are scare, and to this end, we optimized simple checklists for two widely used picture narratives to assess PPA subtypes, PSP and CBS. We employed a hierarchical structure in our

LASSO analysis given the nature of word usage across patient groups. The LASSO models could not differentiate svPPA versus lvPPA, nfvPPA versus PSP and nfvPPA versus PSP and CBS with the target words alone. Supplementing the checklist with MLSE and ACE-R subtest scores improved the differentiation between these groups with excellent within group 4-fold cross-validation accuracies. Out-of-sample test accuracy was also found to be high for svPPA versus lvPPA, which emphasizes the need for further specialist assessments for aphasic groups that cluster based on shared clinical features (i.e. anomia in svPPA and lvPPA, motor speech and/or agrammatism in nfvPPA, PSP and CBS).

Clinical tools that are fast, simple and sensitive to aphasia subtypes including various checklists have previously been proposed for post-stroke aphasia,[64] but to our knowledge, this is the first study to provide a direct comparison of word usage across PPA subtypes and Parkinson-plus disorders and optimize a checklist for these patient groups. Future studies with connected speech samples could employ similar methodologies such as our LASSO models to generate specific word checklists for other picture description tasks, different languages and/or diverse patient groups. The present study could also potentially inform the design of future studies in developing targeted pictures that contain the key vocabulary items that help to differentiate specific clinical groups.

Limitations and clinical implications

There are limitations to our study. We only present clinical, not pathological, diagnoses, although clinic pathological correlations are high for PPA and PSP. Our sample size for the out-of-sample test validation was small particularly for certain groups such as PSP. However, we mitigated the potential limitations of small sample *k*-fold cross-validation by conducting predictive validity testing on an unseen data set. This supports generalizability of our models and word checklists. Future work is warranted to test the generalizability of the word checklists to larger patient samples that span various disease stages, non-English languages and varying levels of demographics including education and geographical regions.

A major aim of the present study was to ameliorate the problem of connected speech analyses being time-consuming, effortful and inconsistent across clinicians and different clinical/research settings. As a result, our systematic analysis of connected speech did not include other acoustic and articulatory measures investigated in prior studies. There are undoubtedly other features of the patients' connected speech that can help with differentiation[43] which are not captured in our approach, but these require expertise and time-consuming transcription and analyses. Finally, we acknowledge that our imaging analyses were exploratory but nonetheless add to the current literature pertaining to regions engaged in connected speech.

In conclusion, we propose that screening for language deficits in PPA and 'motor' disorders like PSP and CBS is achievable with a 1-min sample of connected speech. By focusing on the number and lexico-semantic metrics of the given words, rather than acoustic features, this method is likely to be robust to detect dysarthrophonia from disease, even with reduced bandwidth from remote recordings. The screening test is not a substitute for in-depth neuropsychological assessment, but a contributing tool towards diagnosis. Furthermore, it has the advantage of applicability in resource-limited settings and with limited expertise. Future versions of the test for non-English speakers would further increase the international utility of this approach.

Supplementary material

Supplementary material is available at *Brain Communications* online.

Acknowledgements

We thank our patients and their families for supporting this work.

Funding

This work and the corresponding author (S.K.H.) were supported and funded by the Bill & Melinda Gates Foundation, Seattle, WA, and Gates Cambridge Trust (grant number: OPP1144). This study was supported by the Cambridge Centre for Parkinson-Plus; the Medical Research Council (MRC) (MC_UU_00030/14, MR/P01271X/1 and MR/T033371/1); the Wellcome Trust (220258); the National Institute for Health and Care Research (NIHR) Cambridge Clinical Research Facility and the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014 and NIHR203312); Medical Research Council programme intramural funding (MC_UU_00005/18) and Career Development Award (MR/V031481/1). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR, or the Department of Health and Social Care.

Competing interests

The authors report no competing interests.

Data availability

Unthresholded statistical parametric images are available freely on request of the corresponding or senior author. Word lists produced by participants are available on request. Participant MRI scans and transcripts may be available,

subject to a data sharing agreement required to protect confidentiality and adhere to consent terms.

References

1. Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: A review. *Front Psychol*. 2017;8:269.
2. 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.
3. Fromm D, Greenhouse J, Pudil M, Shi Y, MacWhinney B. Enhancing the classification of aphasia: A statistical analysis using connected speech. *Aphasiology*. 2022;36(12):1492-1519.
4. Bird H, Lambon Ralph MA, Patterson K, Hodges JR. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain Lang*. 2000;73(1):17-49.
5. Hoffman P, Meteyard L, Patterson K. Broadly speaking: Vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*. 2014;55:30-42.
6. Haley KL, Jacks A, Jarrett J, *et al*. Speech metrics and samples that differentiate between nonfluent/agrammatic and logopenic variants of primary progressive aphasia. *J Speech Lang Hear Res*. 2021;64(3):754-775.
7. Fraser KC, Meltzer JA, Graham NL, *et al*. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*. 2014;55:43-60.
8. Wilson SM, Henry ML, Besbris M, *et al*. Connected speech production in three variants of primary progressive aphasia. *Brain*. 2010;133(Pt 7):2069-2088.
9. Peterson KA, Patterson K, Rowe JB. Language impairment in progressive supranuclear palsy and corticobasal syndrome. *J Neurol*. 2021;268(3):796-809.
10. Burrell JR, Hodges JR, Rowe JB. Cognition in corticobasal syndrome and progressive supranuclear palsy: A review. *Mov Disord*. 2014;29(5):684-693.
11. Peterson KA, Jones PS, Patel N, *et al*. Language disorder in progressive supranuclear palsy and corticobasal syndrome: Neural correlates and detection by the MLSE screening tool. *Front Aging Neurosci*. 2021;13:675739.
12. Parjane N, Cho S, Ash S, *et al*. Digital speech analysis in progressive supranuclear palsy and corticobasal syndromes. *J Alzheimers Dis*. 2021;82(1):33-45.
13. Esmonde T, Giles E, Xuereb J, Hodges J. Progressive supranuclear palsy presenting with dynamic aphasia. *J Neurol Neurosurg Psychiatry*. 1996;60(4):403-410.
14. Robinson GA, Spooner D, Harrison WJ. Frontal dynamic aphasia in progressive supranuclear palsy: Distinguishing between generation and fluent sequencing of novel thoughts. *Neuropsychologia*. 2015;77:62-75.
15. Robinson G, Shallice T, Cipolotti L. Dynamic aphasia in progressive supranuclear palsy: A deficit in generating a fluent sequence of novel thought. *Neuropsychologia*. 2006;44(8):1344-1360.
16. Catricalà E, Boschi V, Cuoco S, *et al*. The language profile of progressive supranuclear palsy. *Cortex*. 2019;115:294-308.
17. Del Prete E, Tommasini L, Mazzucchi S, *et al*. Connected speech in progressive supranuclear palsy: A possible role in differential diagnosis. *Neurol Sci*. 2021;42(4):1483-1490.
18. Gross RG, Ash S, McMillan CT, *et al*. Impaired information integration contributes to communication difficulty in corticobasal syndrome. *Cogn Behav Neurol*. 2010;23(1):1-7.
19. de Almeida IJ, Silagi ML, Carthery-Goulart MT, *et al*. The discourse profile in corticobasal syndrome: A comprehensive clinical and biomarker approach. *Brain Sci*. 2022;12(12):1705.
20. Patel N, Peterson KA, Ingram RU, *et al*. A 'Mini linguistic state examination' to classify primary progressive aphasia. *Brain Commun*. 2022;4(2):fcab299.
21. Gorno-Tempini ML, Hillis AE, Weintraub S, *et al*. Classification of primary progressive aphasia and its variants. *Neurology*. 2011;76(11):1006-1014.
22. Hoglinger GU, Respondek G, Stamelou M, *et al*. Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Mov Disord*. 2017;32(6):853-864.
23. Armstrong MJ, Litvan I, Lang AE, *et al*. Criteria for the diagnosis of corticobasal degeneration. *Neurology*. 2013;80(5):496-503.
24. Goodglass HKE. *The assessment of aphasia and related disorders*. Lea & Febiger; 1983.
25. Zimmerer VC, Wibrow M, Varley RA. Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *J Alzheimers Dis*. 2016;53(3):1145-1160.
26. Manning BL, Harpole A, Harriott EM, Postolowicz K, Norton ES. Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. *J Speech Lang Hear Res*. 2020;63(12):3982-3990.
27. Zimmerer VC, Hardy CJD, Eastman J, *et al*. Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: An approach based on usage-frequency. *Cortex*. 2020;133:103-119.
28. Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav Res Methods Instrum Comput*. 1996;28(2):203-208.
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.
30. Shaoul C, Westbury C. Exploring lexical co-occurrence space using HiDEx. *Behav Res Methods*. 2010;42(2):393-413.
31. Brysbaert M, Warriner AB, Kuperman V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods*. 2014;46(3):904-911.
32. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. Age-of-acquisition ratings for 30,000 English words. *Behav Res Methods*. 2012;44(4):978-990.
33. Yarkoni T, Balota D, Yap M. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychon Bull Rev*. 2008;15(5):971-979.
34. Balota DA, Yap MJ, Cortese MJ, *et al*. The English Lexicon Project. *Behav Res Methods*. 2007;39(3):445-459.
35. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage*. 2007;38(1):95-113.
36. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267-288.
37. Crutch SJ, Warrington EK. The influence of refractoriness upon comprehension of non-verbal auditory stimuli. *Neurocase*. 2008;14(6):494-507.
38. Jefferies E, Patterson K, Jones RW, Lambon Ralph MA. Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology*. 2009;23(4):492-499.
39. Henderson SK, Peterson KA, Patterson K, Lambon Ralph MA, Rowe JB. Verbal fluency tests assess global cognitive status but have limited diagnostic differentiation: Evidence from a large-scale examination of six neurodegenerative diseases. *Brain Commun*. 2023;5(2):fcad042.
40. Cho S, Nevler N, Ash S, *et al*. Automated analysis of lexical features in frontotemporal degeneration. *Cortex*. 2021;137:215-231.
41. Cordella C, Dickerson BC, Quimby M, Yunusova Y, Green JR. Slowed articulation rate is a sensitive diagnostic marker for identifying non-fluent primary progressive aphasia. *Aphasiology*. 2017;31(2):241-260.
42. Themistocleous C, Webster K, Afthinos A, Tsapkini K. Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *Am J Speech Lang Pathol*. 2021;30(1S):466-480.
43. Faroqi-Shah Y, Treanor A, Ratner NB, Ficek B, Webster K, Tsapkini K. Using narratives in differential diagnosis of neurodegenerative syndromes. *J Commun Disord*. 2020;85:105994.

44. Garcia AM, Welch AE, Mandelli ML, *et al.* Automated detection of speech timing alterations in autopsy-confirmed nonfluent/agrammatic variant primary progressive aphasia. *Neurology*. 2022; 99(5):e500-e511.
45. Magdalinou NK, Golden HL, Nicholas JM, *et al.* Verbal adynamia in parkinsonian syndromes: Behavioral correlates and neuroanatomical substrate. *Neurocase*. 2018;24(4):204-212.
46. Burrell JR, Ballard KJ, Halliday GM, Hodges JR. Aphasia in progressive supranuclear palsy: As severe as progressive non-fluent aphasia. *J Alzheimers Dis*. 2018;61(2):705-715.
47. Patterson K, MacDonald MC. Sweet nothings: Narrative speech in semantic dementia. In: *From inkmarks to ideas: Current issues in lexical processing*. 1st ed. Psychology Press; 2006:329-347.
48. Meteyard L, Patterson K. The relation between content and structure in language production: An analysis of speech errors in semantic dementia. *Brain Lang*. 2009;110(3):121-134.
49. Lambon Ralph MA, Graham KS, Ellis AW, Hodges JR. Naming in semantic dementia–what matters? *Neuropsychologia*. 1998;36(8): 775-784.
50. Vonk JMJ, Jonkers R, Hubbard HI, Gorno-Tempini ML, Brickman AM, Obler LK. Semantic and lexical features of words dissimilarly affected by non-fluent, logopenic, and semantic primary progressive aphasia. *J Int Neuropsychol Soc*. 2019;25(10):1011-1022.
51. Cho S, Quilico Cousins KA, Shellikeri S, *et al.* Lexical and acoustic speech features relating to Alzheimer disease pathology. *Neurology*. 2022;99(4):e313-e322.
52. Mandelli ML, Vitali P, Santos M, *et al.* Two insular regions are differentially involved in behavioral variant FTD and nonfluent/agrammatic variant PPA. *Cortex*. 2016;74:149-157.
53. Dronkers NF. A new brain region for coordinating speech articulation. *Nature*. 1996;384(6605):159-161.
54. Miceli G, Turriziani P, Caltagirone C, Capasso R, Tomaiuolo F, Caramazza A. The neural correlates of grammatical gender: An fMRI investigation. *J Cogn Neurosci*. 2002;14(4):618-628.
55. Kielar A, Milman L, Bonakdarpour B, Thompson CK. Neural correlates of covert and overt production of tense and agreement morphology: Evidence from fMRI. *J Neurolinguistics*. 2011;24(2): 183-201.
56. Bonilha L, Hillis AE, Wilmskoetter J, *et al.* Neural structures supporting spontaneous and assisted (entrained) speech fluency. *Brain*. 2019;142(12):3951-3962.
57. Nevler N, Ash S, McMillan C, *et al.* Automated analysis of natural speech in amyotrophic lateral sclerosis spectrum disorders. *Neurology*. 2020;95(12):e1629-e1639.
58. Schönberger E, Heim S, Meffert E, *et al.* The neural correlates of agrammatism: Evidence from aphasic and healthy speakers performing an overt picture description task. *Front Psychol*. 2014; 5:246.
59. Kircher TT, Oh TM, Brammer MJ, McGuire PK. Neural correlates of syntax production in schizophrenia. *Br J Psychiatry*. 2005;186: 209-214.
60. Graves WW, Grabowski TJ, Mehta S, Gordon JK. A neural signature of phonological access: Distinguishing the effects of word frequency from familiarity and length in overt picture naming. *J Cogn Neurosci*. 2007;19(4):617-631.
61. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat Neurosci*. 2009;12(6):718-724.
62. Wilson SM, Isenberg AL, Hickok G. Neural correlates of word production stages delineated by parametric modulation of psycholinguistic variables. *Hum Brain Mapp*. 2009;30(11): 3596-3608.
63. Hodgson VJ, Lambon Ralph MA, Jackson RL. Multiple dimensions underlying the functional organization of the language network. *Neuroimage*. 2021;241:118444.
64. Alyahya RSW, Conroy P, Halai AD, Ralph MAL. An efficient, accurate and clinically-applicable index of content word fluency in aphasia. *Aphasiology*. 2022;36(8):921-939.