ACCEPTED FOR PUBLICATION IN *CORTEX* ON 27/08/20

Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: An approach based on usage-frequency

Vitor C. Zimmerer[a], Chris J.D. Hardy[b], James Eastman[c], Sonali Dutta[d], Leo Varnet[e], Rebecca L. Bond[b], Lucy Russell[b], Jonathan D. Rohrer[b], Jason D. Warren[b], Rosemary A. Varley[a]

[a]University College London, Department of Language and Cognition, Chandler House, 2 Wakefield Steet, London WC1 1PF, United Kingdom

[b]University College London, Dementia Research Centre, Gower Street, London, WC1E 6BT, United Kingdom

[c]Southend University Hospital NHS Foundation Trust, Prittlewell Chase, Westcliff-on-Sea, SS0 0RY, United Kingdom

[d]St Andrew's Healthcare, Pound Ln, North Benfleet, Essex, SS12 9JP, United Kingdom

[e]University College London, Department of Speech, Hearing and Phonetic Sciences, Chandler House, 2 Wakefield Steet, London WC1 1PF, United Kingdom

Corresponding author:

Vitor C. Zimmerer

Chandler House

2 Wakefield Street

London WC1N 1PF

v.zimmerer@ucl.ac.uk

ABSTRACT

Language production provides important markers of neurological health. One feature of impairments of language and cognition, such as those that occur in stroke aphasia or Alzheimer's disease, is an overuse of high frequency, "familiar" expressions. We used computerized analysis to profile narrative speech samples from speakers with variants of frontotemporal dementia (FTD), including subtypes of primary progressive aphasia (PPA). Analysis was performed on language samples from 29 speakers with semantic variant PPA (svPPA), 25 speakers with logopenic variant PPA (lvPPA), 34 speakers with non-fluent variant PPA (nfvPPA), 14 speakers with behavioural variant FTD (bvFTD) and 20 older normal controls (NCs). We used frequency and collocation strength measures to determine use of familiar words and word combinations. We also computed word counts, content word ratio and a combination ratio, a measure of the degree to which the individual produces connected language. All dementia subtypes differed significantly from NCs. The most discriminating variables were word count, combination ratio, and content word ratio, each of which distinguished at least one dementia group from NCs. All participants with PPA, but not participants with bvFTD, produced significantly more frequent forms at the level of content words, word combinations, or both. Each dementia group differed from the others on at least one variable, and language production variables correlated with established behavioral measures of disease progression. A machine learning classifier, using narrative speech variables, achieved 90% accuracy when classifying samples as NC or dementia, and 59.4% accuracy when matching samples to their diagnostic group. Automated quantification of spontaneous speech in both language-led and non-language led dementias, is feasible. It allows extraction of syndromic profiles that complement those derived from standardized tests, warranting further evaluation as candidate biomarkers. Inclusion of frequency-based language variables benefits profiling and classification.

1. Background

In clinical practice, diagnosis and tracking of language in dementia, such as in primary progressive aphasias (PPA), commonly relies on a range of formal neuropsychological tests, such as picture naming or sentence-picture matching, as well as descriptions of spontaneous language output, e.g. as “non-fluent” or “jargon” (Gorno-Tempini et al., 2011; Marshall et al., 2018). However, there is considerable interest in quantification of broader aspects of spontaneous speech, which in turn may support early identification of decline in language function and allow sensitive tracking of behavior change (Ash et al., 2013; Boschi et al., 2017; Fraser et al., 2014; Fraser, Meltzer, & Rudzicz, 2015; Nevler et al., 2017; Wilson et al., 2010; Zimmerer, Wibrow, & Varley, 2016). Investigations of spontaneous production also provide more direct insight into the functional difficulties experienced by patients and their communication partners, and may be less subject to test anxiety in comparison to formal tests (Keady & Gillard, 2002).

Traditionally in studies of language in cognitive disorders, the focus has been on structural properties of language output, such as the complexity of syntactic structures, the distribution of different word classes, and the number of errors. These measures are rooted in formal grammar traditions, such as generative linguistics (e.g. Chomsky, 1981), that focus on word types or classes of grammatical operations. Processing difficulties (and likelihood of impairment in a clinical condition) are seen as a result of disruption of a formal grammatical operation, such as past tense inflection or transformation from canonical word order (see also Avrutin, 2000; Grodzinsky, 2000; Mauner, Fromkin, & Cornell, 1993). However, recent studies reveal that patterns of impairment do not necessarily match to lexical or grammatical category boundaries, and that usage-frequency, i.e., how often an expression is encountered in everyday communication, is an important predictor of its likely resilience to neurological damage (for discussions see Gahl & Menn, 2016; Zimmerer, Dąbrowska, & Varley, under review).

For example, interpretation of passive constructions, often described as impaired in people with non-fluent, Broca’s type aphasia, is less disrupted if the utterance contains a verb that is biased towards passive use, e.g. c*lean* such as in *The room was cleaned by the maid* (Gahl et al., 2003; Menn, Gahl, Holland, Ramsberger, & Jurafsky, 2003). Similarly, plural morphology is more available, and even overproduced, when the noun is biased towards its plural form, e.g. *slippers* (Hatchard & Lieven, 2019). Very familiar expressions (e.g., *I don't know*), frequent function word clusters (e.g., *I can't*), or frequent compounds (e.g., *red cross*) may become entrenched as formulaic expressions and preserved even in the case of severe aphasia (Code, 1982; Mondini, Jarema, Luzzatti, Burani, & Semenza, 2002; Van Lancker Sidtis & Yang, 2016; Zimmerer, Newman, Thomson, Coleman, & Varley, 2018), or Alzheimer's disease (Bridges & Van Lancker Sidtis, 2013; Wray, 2014; Zimmerer et al., 2016). The presence of these phenomena can be traced back to Broca's famous case "Tan", whose production was extremely limited but who retained the expletive *Sacré nom de Dieu* (Code, 2013). Far from being "automatic" or non-propositional speech in a pathological sense, formulas are often used appropriately (Bruns et al., 2019; Van Lancker Sidtis, 2012). Because they are represented as a sequence of strongly associated words, or as a single "word-like" unit, they pose fewer demands on word retrieval and combinatorial systems (Conklin & Schmitt, 2012; Siyanova-Chanturia, Conklin, Caffarra, Kaan, & van Heuven, 2017).

For this reason, we expect that most disruptions of lexical or grammatical production will result in individuals compensating by over-relying on frequent language forms. Fraser et al. (2014) showed that speakers with svPPA and nfvPPA produced more frequent words in spontaneous speech. In the current study, we explored usage frequency of words but also the collocation strength of word combinations (which is based on frequency). We also examined performance in a broader range of FTD sub-groups. We analysed samples of spontaneous speech from three PPA variant syndromes: semantic variant (svPPA), logopenic variant (lvPPA) and non-fluent variant (nfvPPA). We compared language measures with those derived from participants with behavioral variant frontotemporal dementia (bvFTD) as a disease control group without a primary language disorder, and with older normal control individuals (NC). We used the Frequency in Language Analysis Tool (FLAT), an automated script for quantification of language features.

FLAT extracts each word and word combination from a sample and looks up their frequencies in the spoken subcorpus of the British National Corpus (BNC; 2007), a 10-million-word collection of normative samples recorded across a range of communication contexts, geographical regions and demographic groups. It also uses these values to calculate collocation strength, a variable that indicates the degree to which words within a combination are associated with one another. Stronger collocations are more likely to be processed as formulas. FLAT also characterizes grammatical profile via content/function word ratios and combination ratio - a measure of how connected the language output is. In previous work we employed FLAT to characterize language production in stroke aphasia (Bruns et al., 2019; Zimmerer, Newman, Thomson, Coleman, & Varley, 2018) and Alzheimer's disease (Zimmerer et al., 2016). These studies found increased word frequency and collocation strength in both groups. In Alzheimer's disease, the correlation between collocation strength of combinations and estimated time post-disease onset was significant.

Methods

1.1 Hypotheses

We hypothesized that, in pairwise group comparisons, the spontaneous speech of each group would differ from all others on at least one variable. While we had expectations about specific variables and directions, such as that svPPA would produce more common content words than NCs, and that speakers with nfvPPA would produce less connected language, the novelty of the research warrants a more exploratory approach, and we therefore chose bi-directional inferential tests. We also expected that within each group, properties of language production would be related to validated measures of dementia progression.

1.2 Participants and samples

This study employs secondary analysis of data from University College London's (UCL) Dementia Research Centre. Data collection was approved by the University College London institutional ethics committee (reference no. Q6/Q051/52) and all participants gave informed consent in accordance with the Declaration of Helsinki. Participants were recruited via a tertiary cognitive disorders clinic as part of a larger neuropsychological and neuroimaging study of frontotemporal dementia and related disorders. Speech samples were recorded on first assessment. Time points varied relative to initial diagnosis.

The recorded interviews were conducted by psychology research assistants working in the Dementia Research Centre, UCL. All examiners were trained to ensure consistency in conduct of the assessments. Interviews were undertaken in a quiet room, and speech samples were recorded using handheld recording devices for subsequent offline transcription.

Participants were asked to talk about their last holiday and, using prompts if necessary, encouraged to talk for up to three minutes. The procedure was designed to be open-ended, in order to capture the wide range of fluency and general language difficulties experienced by patients. Examples of prompts included, “Where did you go?”, “How long were you there for?”, “How did you get there?”, and “What did you do there?”.

In comparison with picture or video description tasks, holiday narratives are less constrained by topic and lexical items. This lack of constraint can introduce additional heterogeneity into samples, since the participant’s experience or topic choice may influence type of language elicited (e.g., word selection), and participants have more freedom to employ compensatory strategies. However, compared to description tasks open questions elicit more naturalistic linguistic behaviour, increasing ecological validity of the measures.

We selected recordings of spontaneous speech based on the following criteria: (1) diagnosis of one of the three canonical PPA types or bvFTD, or being part of the healthy older control sample; (2) adequate audio quality for transcription since, in a few cases, there was too much background noise or the microphone had been placed too far from the participant. Speech samples were orthographically transcribed from the audio recording. The examiner prompts were removed from transcriptions, which was carried out by J.E. and S.D. under supervision of V.Z. In cases of portions of unclear audio or speech, an experienced clinician (R.A.V.) was consulted and consensus was reached. In cases of phonological errors, the target word was transcribed if it was recognizable on the basis of context and phonological form. If not, it was excluded. Place and person names were also excluded.

The final sample set consisted of 20 NCs, 29 participants with svPPA, 25 participants with lvPPA, 34 participants with nfvPPA and 14 people with bvFTD (see Table 1 for summary). All patients met consensus criteria for their diagnostic group (Gorno-Tempini et al., 2011; Rascovsky et al., 2011). Age did not differ significantly across groups, F(1,120) = .01, p = .937, though pairwise comparisons revealed that individuals with nfvPPA were significantly older than NCs (p = .009).

A range of standardized tests was used to profile cognitive status, although protocols were not the same in all cases. All groups except NCs were assessed using the Mini Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975). Relative to a large UK population sample (Huppert, Cabelli, Matthews, 2005), average MMSE scores for svPPA and lvPPA were below the 5th percentile, nfvPPA scores were at approximately the 25th percentile, and bvFTD scores at approximately the 5th percentile.

Participants were also tested on the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 2011), the Recognition Memory Test (RMT; Warrington, 1984) for words and faces, which assesses episodic memory, maximum forward and reverse digit span (Wechsler, 2011), the Graded Difficulty Arithmetic (GDA; Jackson & Warrington, 1986) which tests calculation, the Graded Naming Test (GNT; McKenna & Warrington, 1983) which tests word retrieval, the British Picture Vocabulary Scale (BPVS; Dunn, Dunn, Whetton, & Burley, 1997) which tests word comprehension, and the Visual Object Space Perception (VOSP; Warrington & James, 1991). See Appendix A for test summaries and group comparisons.

Table 1. Demographic, clinical and general neuropsychological data for the participant groups. Mean (standard deviation) values are shown. Raw scores are presented. Bold numbers indicate significant differences from normal controls (p < .05).

Key: bvFTD, behavioral variant frontotemporal dementia; BPVS, British Picture Vocabulary Scale; Controls, healthy control group; Digit span forward/reverse, maximum digit span recorded; F, Female; GDA, Graded Difficulty Arithmetic; GNT, Graded Naming Test; lvPPA; patient group with logopenic variant primary progressive aphasia; M, Male; MMSE, Mini-Mental State Examination; nfvPPA, patient group with non-fluent variant primary progressive aphasia; RMT, Recognition Memory Test; svPPA, patient group with semantic variant primary progressive aphasia; VOSP, Visual Object Space Perception; WASI, Wechsler Abbreviated Scale of Intelligence. Maximum value in parentheses; n indicates number of scores available for the respective group and test.

	svPPA n = 29	**lvPPA** n = 25	**nfvPPA** n = 34	**bvFTD** n = 14	**Controls** n = 20
Demographic and clinical					
No. (M:F)	17:12	12:13	13:20	**12:2**	10:10
Age (years)	64.00 (7.84)	63.32 (13.61)	**69.82 (8.41)**	64.36 (8.45)	62.78 (7.27)
MMSE (/30)	20.39 (8.19) n = 25	19.47 (7.85) n = 19	26.96 (10.15) n = 23	23.83 (7.33) n = 12	-
General intellect (WASI)	n = 25	n = 20	n = 22	n = 14	n = 20
WASI Vocabulary (/80)	**24.07 (22.38)**	**19.00 (16.78)**	**32.59 (20.12)**	**39.64 (25.18)**	70.40 (4.71)
WASI Block design (/71)	**32.04 (18.85)**	**14.95 (16.94)**	**23.05 (18.53)**	**20.93 (17.91)**	45.65 (10.78)
WASI Similarities (/48)	**14.36 (11.31)**	**14.05 (13.60)**	**20.05 (13.92)**	**19.50 (12.25)**	39.95 (4.73)
WASI Matrix Reasoning (/35)	**18.32 (8.86)**	**11.35 (9.20)**	**15.95 (9.17)**	**13.57 (8.34)**	25.25 (1.8)
Episodic memory					
RMT Words (/50)	**34.19 (7.59)** n = 21	**32.31 (9.60)** n = 16	**43.47 (6.01)** n = 19	**35.50 (7.74)** n = 12	48.40 (1.82) n = 20
RMT Faces (/50)	**34.86 (8.14)** n = 22	**34.32 (7.20)** n = 19	**38.35 (6.28)** n = 20	**33.92 (6.20)** n = 13	43.10 (5.06) n = 20
Working memory					

Digit span forward (max)	6.40 (1.63) n = 25	**3.45 (1.28)** n = 20	**5.05 (1.57)** n = 20	6.15 (1.41) n = 13	6.85 (1.27) n = 20
Executive function					
Digit span reverse (max)	4.72 (1.99) n = 25	**2.70 (0.98)** n =20	**3.56 (1.76)** n = 18	**3.77 (2.01)** n = 13	5.20 (1.06) n = 20
Posterior cortical skills					
GDA Calculation (/24)[j]	**9.54 (7.94)** n = 24	**2.35 (2.74)** n = 17	**5.17 (5.45)** n = 18	**10.25 (7.31)** n = 12	16.15 (3.76) n = 20
VOSP Object Decision (/20)[k]	**16.40 (3.34)** n = 25	17.0 (3.34) n = 20	**16.19 (3.89)** n = 21	**15.29 (2.95)** n = 14	18.70 (1.42) n = 20
Word retrieval					
GNT (/30)[l]	**1.21 (2.87)** n = 24	**4.75 (5.97)** n = 20	**13.90 (9.10)** n = 21	**10.93 (7.49)** n = 14	25.10 (3.41) n = 20
Word comprehension					
BPVS (/150)[m]	**69.10 (55.47)** n = 24	**125.37 (26.56)** n = 19	128.90 (35.70) n = 21	128.43 (34.16) n = 14	147.50 (2.01) n = 20

1.1 Sample analysis procedure

Samples were formatted for analysis with the Language Analysis Tool (FLAT) (Bruns et al., 2019; Zimmerer et al., 2018; Zimmerer et al., 2016). See the online supplement for annotation rules and examples. The FLAT is a computer program that works with any text, e.g. transcribed speech. It extracts from the sample every word, bigram (two-word combination) and trigram (three-word combination), moving through the text one word at a time. For example, the expression *The man eats bread* consists of four words, three bigrams (*the man*; *man eats*; *eats bread*) and two trigrams (*the man eats*; *man eats bread*). It classifies each word as a content word (e.g., *house*, *climb*, *fast*) or function word (e.g., *it*, *the*, *when*), and calculates the proportion of content words. It also calculates combination ratio, a measure of fluency, by dividing the number of trigrams in the corresponding sample by the number of words. For example, a sample consisting of the utterance *My parents came to my house and we all had dinner* together contains 12 words and 10 trigrams (*my parents came, parents came to, came to my*, etc.). The combination ratio of would be .83 (10 divided by 12). In contrast, *My parents. They came to my house. We all had dinner together* contains 12 words and 6 trigrams (*They came to, to my house, we all had*, etc.). That sample would have a combination ratio of .5.

Importantly, the FLAT looks up usage frequency of each word, bigram or trigram (reported in instances per million words) in the spoken subsection of the British National Corpus (BNC; 2007). At the single word level, we were interested in the frequency of each content word in the sample and averaged these values for each participant. The most novel contribution of FLAT however is the inclusion of collocation strength of word combinations in the samples. Collocation strength measures the frequency with which words co-occur relative to their “expected frequency”, i.e., if word order was random. The value can be high even if the frequency of individual words is low, e.g., the combination *plate tectonics* has high collocation strength despite containing low frequency words. Therefore, while individual word frequency reflects the difficulty of retrieving word forms (more frequent is considered easier), collocation strength reflects the difficulty posed in combining individual words (higher collocation strength is considered easier). We focused on bigrams since in previous studies, these yielded bigger effect sizes than trigrams when comparing participants with neurological damage to controls (Bruns et al., 2019; Zimmerer et al., 2016).

We used t-scores as our measure of collocation strength.

The t-scores formula for a bigram collocation of words *a* and *b* is

$$t\text{-}score_{ab} = \frac{frequency_{ab} - expected\ frequency_{ab}}{\sqrt{frequency_{ab}}}$$

where $frequency_{ab}$ stands for the observed frequency of the bigram in the spoken BNC. In the 10 million word corpus, the expected frequency for a bigram *ab* is

$$expected\ frequency_{ab} = \frac{frequency_{a}\, frequency_{b}}{10{,}000{,}000}$$

where $frequency_{a}$ and $frequency_{b}$ stand for the observed frequency of the individual words of the bigram in the spoken BNC. Compared to other collocation strength measures such as Mutual Information, t-scores are better suited for lower frequency combinations (Church & Hanks, 1990; Gries, 2010). Collocation strength averages excluded bigrams with a frequency of zero. Table 2 summarizes variables produced by FLAT.

Table 2. Properties of language production measured by the automated Frequency in Language Analysis Tool (FLAT).

Variable name	**Description**	**Type of marker**
Word count	Measure of sample size.	Quantity of language output; verbal responsiveness.
Combination ratio $(\frac{trigram\ count}{word\ count})$	Measure of connected language, i.e., the degree to which the speaker produces longer combinations as opposed to one- and two-word fragments.	Verbal responsiveness; sentence complexity.
Content word ratio $(\frac{content\ word\ count}{word\ count})$	Proportion of content words (vs. function words) in a language sample.	Relationship between lexical and grammatical capacity. Can indicate lexical impairment (too few content words) or grammatical impairment (too few function words).
Content word frequency	Average usage frequency of content words, measured in occurrences per million words in the spoken BNC.	Lexical capacity. Over-representation of more frequent words can indicate lexical impairment.
Collocation strength (bigram t-score) $\frac{frequency\ - expected\ frequency}{\sqrt{frequency}}$	The degree to which words in a combination appear more often together than	Capacity to produce new or novel utterances, e.g., production of more highly collocated utterances indicates

	would be expected by chance.	stronger reliance on formulaic language.

As a first step, we looked for differences between groups for each variable, treating each as a separate hypothesis. For word count, combination ratio and content word ratio, there is one data point for each participant, and we computed analyses of covariance with participant group as the independent variable, age as covariate and the language measure as dependent variable. For content word frequency and collocation strength, there were multiple data points for each participant. We exploited this through use of *R* (R Core Team, 2019) and *lme4* (Bates, Mächler, Bolker, & Walker, 2015) to perform linear mixed effects analyses of the relationship between group and the respective variables. As fixed effects, we entered group and age (without interaction term) into the model. We entered participant ID ("Transcript") as a random effect. Adding an additional random slope (Age|Group) resulted in overfitting (singular fits). Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. We obtained p-values via likelihood ratio tests of the full model (DV ~ Age + Group + (1|Transcript)) against the model without the effect of group (DV ~ Age + (1|Transcript)). We adjusted significance thresholds for between-group comparisons.

We then explored correlations between FLAT variables and validated markers of disease within each group by examining Pearson's r correlations between language measures and standardized test scores. Our primary question was to determine the properties of spontaneous language most strongly related with standardized verbal and non-verbal cognitive measures, as well as with MMSE scores. We had no specific hypotheses, beyond the expectation that deviations from the control group in the observed direction (lower word count, fewer combinations, higher frequency and collocation strength) would be associated with poorer performance. We built correlation heatmaps to visualize these relationships using the ggplot2 package for R (R Development Core Team, 2008; Wickham et al., 2016).

While group comparisons reveal differences with regard to specific language features, the combination of features into the profiles associated with a type of dementia have the greatest potential in supporting clinical diagnosis. Patterns of variable distributions may be specific enough to a group that, even in the case where group effects on single variables are not significant, the variables can still contribute to classification. In the final step of the analysis, we used a machine learning classifier to determine the degree to which the combination of FLAT variables can distinguish between diagnosis groups (see 2.3). We chose a support vector machine (SVM), which has been established as a text classifier that can use multiple variables (Joachims, 1998; Meyer, Leisch, & Hornik, 2003). We used the SVM function of MATLAB's Classification Learner app ("MATLAB and Statistics and Machine Learning Toolbox," 2018).

Original data and test materials will be shared in secure, anonymised form (to protect patient confidentiality) with researchers based at other academic institutions, following an email request to the corresponding author. This may include a material transfer agreement to cover data exchange between the relevant institutions.

2. Results

2.1 Group comparisons

Table 3 displays group averages and main effects for between-group comparisons with age entered as covariate. All independent variables yielded significant main effects in between-group comparisons. The biggest effect size was for combination ratio, followed by content word frequency, word count, content word ratio and collocation strength. Figure 1 visualizes the language profiles of the dementia subtypes in relation to NCs.

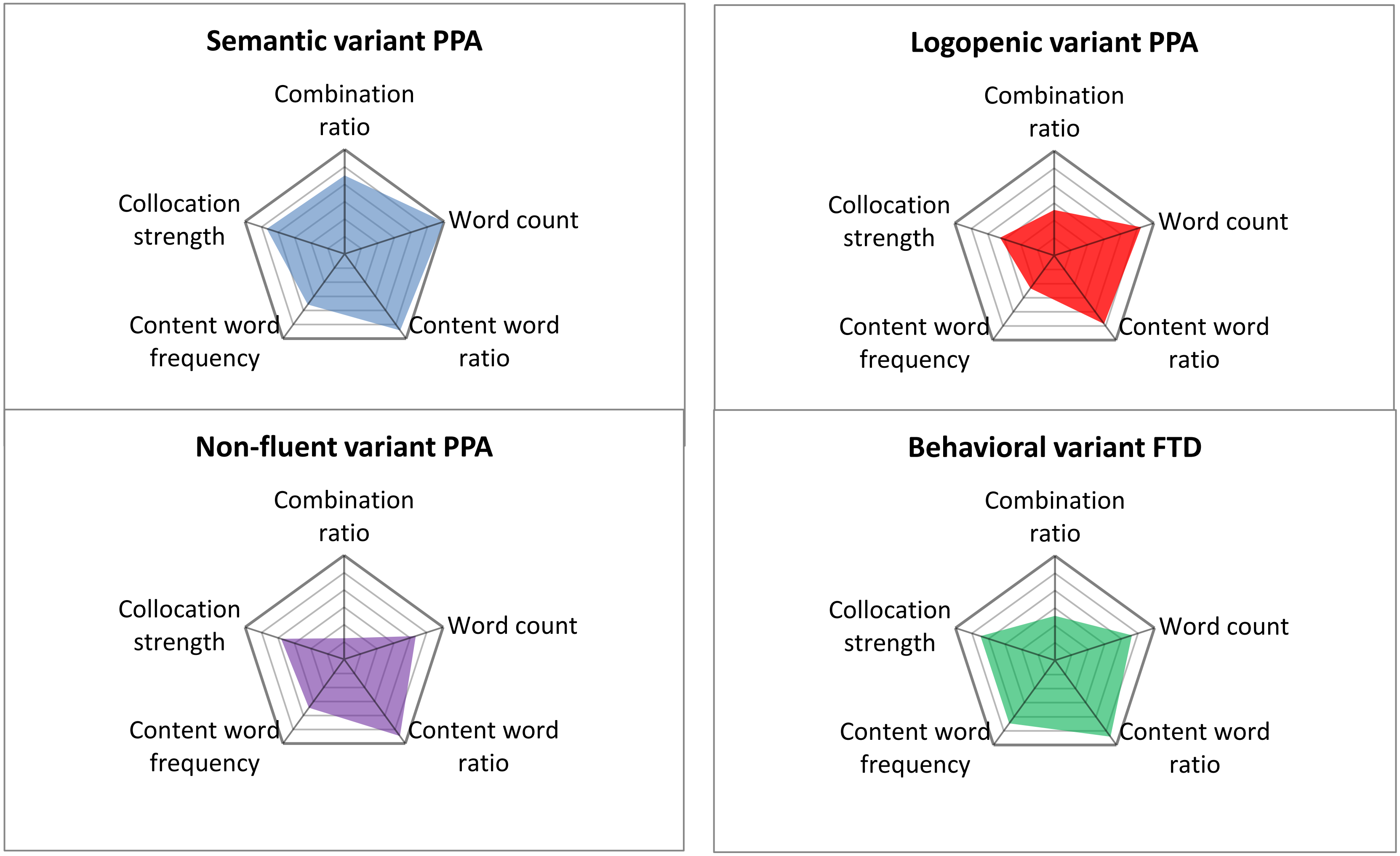
For pairwise comparisons, we used Bonferroni adjustments of significance threshold based on ten comparisons for each variable (each group with each other group). The adjusted threshold was $p < .005$. For mean differences, confidence intervals and p values, see Appendix B. Compared to NCs, speakers with svPPA produced more frequent content words and stronger collocations. Speakers with lvPPA produced language that was less connected, contained fewer and more frequent content words, and stronger collocations. Speakers with nfvPPA produced fewer words, less connected language, more frequent content words and stronger collocations. Speakers with bvFTD produced fewer words, less connected output and more frequent content words. Each dementia subgroup differed from each other subgroup on at least one variable (Appendix B).

Table 3. Group averages (SD) and main effects with age as covariate. ANCOVAs were used for word count, combination ratio and content word ratio. For content word frequency and collocation strength, we compared linear mixed effect models which included age as predictor with models which included age and group. A significant difference indicated that group was a significant predictor. We indicate statistical significance of pairwise differences between groups in each cell: * = sig. different from controls; ^{S} = sig. different from svPPA; ^{L} = sig. different from lvPPA; ^{N} = sig. different from nfvPPA; ^{B} = sig. different from bvFTD. A single character denotes p < .05 (n.s. after adjustment for multiple comparisons); two characters denote p < .005 (sig. after adjustment); three characters denote p < .001. For complete inferential statistics, see Appendix B.

Variable	**Normal controls**	**Semantic variant PPA**	**Logopenic variant PPA**	**Non-fluent variant PPA**	**Behavioral variant FTD**	**Main effect (Age as covariate)**
Word count	260.7 (169)	248.7^{LNNNBBB} (141.3)	156.1*^{SN} (98.0)	87.2***^{SSSL} (80.7)	166.2***^{SSS} (134.6)	F(4,116) = 9.813, p < .001, η^2 = .25
Combination ratio	.77 (.05)	.69*^{LLLNNNB} (.07)	.60***^{SSSNNN} (.09)	.43***^{SSSLLLBBB} (.18)	.60***^{SNNN} (.09)	F(4,116) = 36.717, p < .001, η^2 = .56
Content word ratio	.37 (.04)	.33*^{LL} (.04)	.28***^{SSNNB} (.05)	.35^{LL} (.09)	.33*^{L} (.07)	F(4,116) = 5.878, p < .001, η^2 = .17

Content word frequency (per million)	524 (185)	969***^{LLLNNN} (337)	1278***^{SSSB} (461)	966***^{NNNB} (489)	820***^{LN} (382)	β = 4608, SE = 903, χ^2(1) = 18.949, p < .001
Bigram collocation strength (t-scores)	24.51 (4.07)	29.60***^{LLNN} (4.00)	35.32***^{SSB} (11.90)	32.29***^{SLB} (12.27)	30.37*^{LN} (10.92)	β = 1.74, SE = .44, χ^2(1) = 13.95, p < .001

Figure 1. Radar plot visualization of language profiles of different dementia groups. Data were residualized over participant age in order to account for age differences, and then normalized using control means and standard deviations. The outer line in each plot represents the control mean; each line towards the center represents a distance of one standard deviation from the control mean.



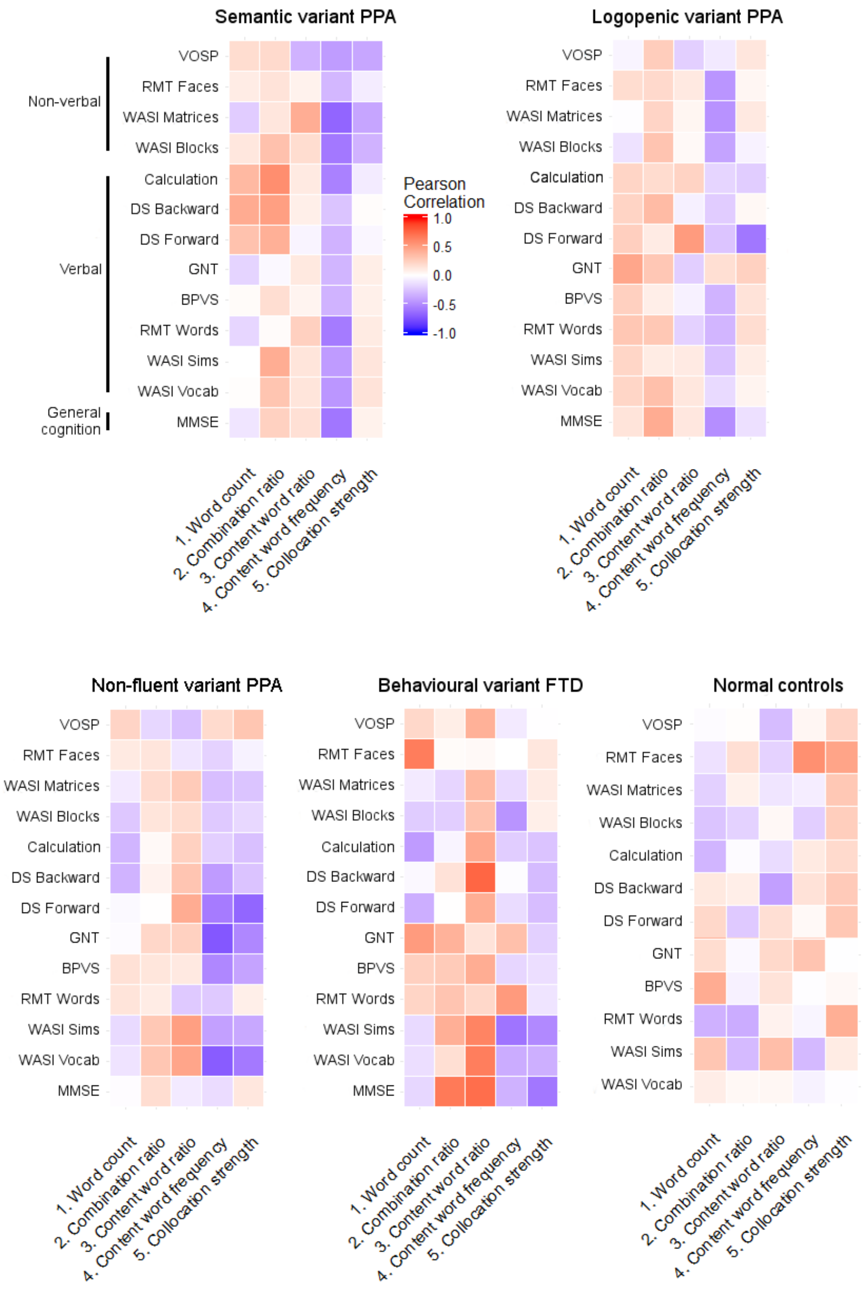
2.2 Relationship with standardized test scores

We expected the deviations from normative language production observed in 2.1 (lower number of words, lower combination ratio, lower content word ratios, higher content word frequency, greater collocation strength) to be associated with worse performance in standardized tests of language and cognition. However, we had no specific hypotheses about how individual FLAT variables would correlate with individual test scores. This part of the analysis is therefore more exploratory. We started with a correlation heatmap (Figure 2), which suggests that in some participant groups, clusters of test scores correlated with FLAT variables. One meaningful categorization of tests is whether they are predominantly verbal (naming tests, verbal memory tests, verbal calculation) or non-verbal (object and space perception, face recognition, matrix- and block reasoning). We calculated composite variables for verbal and non-verbal tests (see Figure 2 for categorization) by first transforming raw scores into z-scores based on control means and standard deviations, and then averaging all scores grouped under one category. We treated MMSE scores separately as a measure of general cognition. These were not available for controls because the MMSE was not part of their protocol. For each correlation, we included participants for which all test scores that made up the composite variable were available. We report correlations with p < .05.

Results are consistent with the visualization in Figure 2. In participants with svPPA, poorer non-verbal test performance was associated with higher content word frequency, r(20) = -.720, p < .001, and higher bigram t-scores, r(20) = -.501, p = .024. Lower verbal test performance was associated with higher content word frequency, r(19) = -.667, p = .002, as were lower MMSE scores, r(23) = -.593, p = .003. In participants with lvPPA, lower non-verbal test scores were associated with higher content word frequency, r(19) = -.517, p = .013. Lower MMSE scores were also associated with higher content word frequency, r(18) = -.488, p = .04. In the nfPPA group, none of the correlations between FLAT variables and composite variables or MMSE scores were significant below the threshold of p < .05. In people with bvFTD, poorer MMSE performance was associated with lower proportions of content words, r(12) = .725, p = .008, lower combination ratio, r(12) = .673, p = .016, and lower bigram t-scores, r(12) = -.587, p = .045. In controls, bigram t-scores correlated with non-verbal scores, r(20) = .461, p = .041.

Note that analysis using individual tests, rather than composite scores, showed some notable relationships, including in the nfPPA group. We report correlations with individual tests with p < .05 in Appendix C.

Figure 2. Correlation heatmaps displaying the relationship between spontaneous language output and standardized measures of language and cognition. Colors indicate effect size (Pearson's r). Tests were categorized according to whether administration and stimuli are predominantly verbal or non-verbal. The MMSE was categorized as a test of general cognition, and was not administered to controls.


Semantic variant PPA

Logopenic variant PPA

Non-fluent variant PPA

Behavioural variant FTD

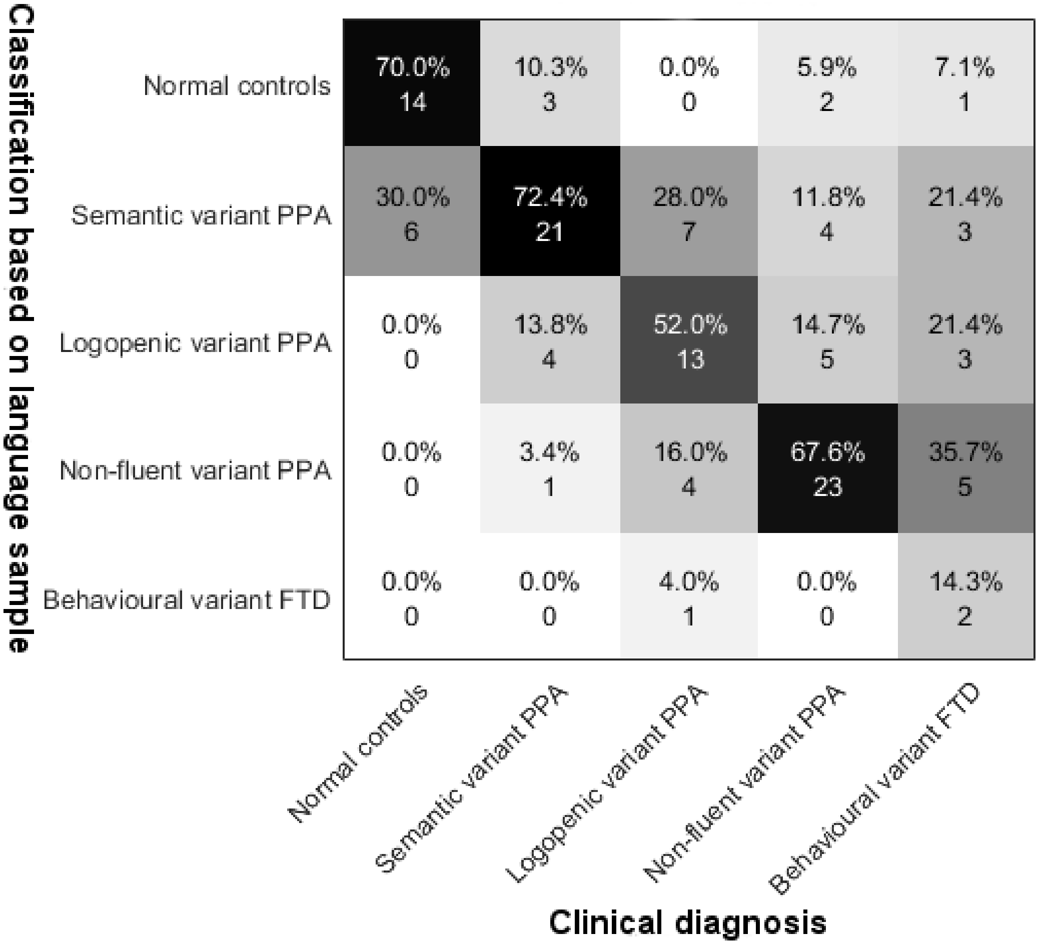
Normal controls

2.3 Machine learning classification

In a final, post-hoc analysis we applied machine learning methods to determine how FLAT data can be used to categorize participants into the five participant groups based solely on their narrative samples. We used a SVM, a complementary linear classifier approach to compare the five groups based on the results of the FLAT analysis. The SVM classifier was trained on a subset of data (80%, randomly selected) to categorize each individual as belonging to one of the five groups based on a linear combination of the five variables (word count, combination ratio, content word ratio, content word frequency, collocation strength). The prediction accuracy of the obtained classifier was then evaluated on the remaining 20% of the data set (test set), in terms of the percentage of correct prediction, and the classification matrix. This procedure was then repeated five times so that each participant was in the test set once. The accuracy of the five tests was averaged to determine overall accuracy of the model.

Against a chance level of 20%, the SVM classifier achieved a prediction accuracy of 59.8%. The classification matrix for the model prediction is shown in Figure 3. The model was most successful in identifying NCs (70% correct classification), speakers with svPPA (72.4% correct) and speakers with nfvPPA (67.6% correct). The model's performance was less accurate for lvPPA (52% correct), and strikingly inaccurate for bvFTD (14.3% correct), with 78.5% of bvFTD samples being classified as one of the PPA types. Of all NCs, 30% (six participants) were classified as svPPA. Six participants with dementia (10.3% of speakers with svPPA, 5.9% of speakers with nfvPPA, 7.1% of speakers with bvFTD) were classified as NCs. The classification matrix is therefore in line with the group comparisons which showed that speakers with svPPA were the group most similar to NCs. It also illustrates the large overlap between language features in bvFTD, as measured by FLAT, and the features of PPAs. If regarded as a detector of dementia in general (two categories: NCs vs. pooled dementia groups), accuracy increases to 90%, with a true positive rate of 94% and a false positive rate of 30%.

Figure 3. Classification matrix based on the SVM classifier. Columns display the clinical diagnosis. Rows show the predicted class on the basis of the five FLAT variables. Percentages represent the proportion of members of the true class with a given diagnosis. Beneath is a raw count of the same members.



	Normal controls	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioural variant FTD
Normal controls	70.0% 14	10.3% 3	0.0% 0	5.9% 2	7.1% 1
Semantic variant PPA	30.0% 6	72.4% 21	28.0% 7	11.8% 4	21.4% 3
Logopenic variant PPA	0.0% 0	13.8% 4	52.0% 13	14.7% 5	21.4% 3
Non-fluent variant PPA	0.0% 0	3.4% 1	16.0% 4	67.6% 23	35.7% 5
Behavioural variant FTD	0.0% 0	0.0% 0	4.0% 1	0.0% 0	14.3% 2

3. Discussion

We ran an automated analysis of spontaneous language production in healthy speakers, canonical PPAs and bvFTD. Group comparisons showed significant differences between all dementia groups and healthy speakers as well as between each dementia subtype. In all dementia groups, there were correlations between language production variables and validated behavioral measures of disease progression. A machine learning classifier, using the output from the analysis, categorized samples with a success rate three times better than chance.

Current diagnostic guidelines mostly limit their description of spontaneous production to reduction of output, grammatical simplification or errors (Gorno-Tempini et al., 2011; Marshall et al., 2018; Wilson et al., 2010). There is strong evidence for another type of simplification, namely increased use of common words and/or phrases. Previous research showed that speakers with svPPA and, to a lesser degree, speakers with nfvPPA, tend to produce more common words (Fraser et al., 2014). Our data support these conclusions and show that this pattern also exists in lvPPA. We also saw differences in the ratio of content words. Speakers with svPPA had a lower content word ratio, consistent with reports of higher pronoun use as a sign of semantic difficulties (e.g. "it" instead of a concrete noun) (Fraser et al., 2014). Speakers with lvPPA also displayed a lower ratio, likely a result of lexical impairment attributed to this population.

Moreover, we show that speakers with nfPPA and lvPPA used more familiar word combinations, as indicated by collocation strength. According to usage-based language theories such as Construction Grammar (Goldberg, 2003), familiar combinations are not produced by retrieving each individual word and combining words using abstract grammatical rules, but through access to holistic representations of the entire phrase or utterance. In many cases, word combinations may be represented as a single "word". Holistic processing reduces demands on language networks both at the lexical level, as fewer units need to be retrieved, and at the grammatical level, as fewer units need to be combined. Both lexical and grammatical impairment should therefore result in an overuse of highly collocated combinations. Our data suggest that this is the case for nfvPPA, primarily a grammatical impairment, and lvPPA, a primarily lexical impairment. Previous studies using FLAT have shown that this is also the case for Alzheimer's disease and both Broca's type and Wernicke's type aphasia (Zimmerer et al., 2018; Zimmerer et al., 2016). We regard them as one aspect of "formulaic language" (Wray, 2012), i.e. the set of combinations, phrases and clauses which consist of words which are strongly associated with each other either because of frequent cooccurrence (measured by collocation strength in this study) or because of idiomatic meaning or use (Bridges & Van Lancker Sidtis, 2013). Overuse of formulaic language forms can have a substantial impact on communication, and therefore quality of life, as the speaker will find it difficult to adapt to unfamiliar situations and speak about new thoughts and needs. There is also the risk of locking conversations into a small number of repeating discourse patterns (Wray, 2011). This is one of the ways in which language measures not only indicate cognitive change, but provide insight into difficulties with social participation.

Formulaic language can also explain why differences in content/function word ratio in spontaneous speech were relatively small even between non-fluent speakers and NCs. Many strongly collocated combinations contain function word clusters, such as *I don't* in the expression *I don't know*. They comprise a substantial proportion of aphasic production (Zimmerer et al., 2018). While one striking feature of non-fluent aphasia is omission of function words, resulting in a "telegraphic" style, formula overuse can result in non-fluent speakers producing a similar ratio of content words to neurotypical speakers over the entire sample, as is the case in the current study.

Our data also support earlier studies which found that bvFTD is associated with significant changes in linguistic behavior (Hardy et al., 2015). In our bvFTD samples, language change affected less the quality of words and word combinations, but rather their quantity, as speakers spoke less and produced shorter utterances. This reduction was roughly on a par with speakers with lvPPA, but while in the latter group this likely reflects impairment of linguistic representations, the linguistic profile in bvFTD may be explained by changes to mood and social behavior, as well as impairment of executive function. However, the differences between bvFTD and NC groups in content word ratio and usage frequency were significant before correction for multiple comparisons (Appendix B), and these findings are consistent with results from studies of word naming and comprehension in bvFTD that also revealed lexical disruption (Hardy et al., 2015). Similarly, other differences that were close to significance thresholds, such as lower combination ratio in speakers with svPPA, may prove important for modelling language in dementia.

Correlations with validated cognitive measures of dementia progression provide further evidence for the clinical relevance of new variables. Properties of spontaneous language samples were linked to performance on standardized language and cognitive tests, although the precise nature of these relationships differed from profile to profile. Generally, a decrease in fluency, and production of more frequent words and strongly collocated combinations were associated with poorer performance in standardized tests. Relationships were strongest in svPPA, lvPPA and bvFTD. Effect sizes were particularly large in the svPPA sample, supporting accounts that regard svPPA not merely as a language impairment, but broader general breakdown of cognitive capacity as multimodal conceptual systems become affected (Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges, 2000; Gorno-Tempini et al., 2011). The data also underline that changes in linguistic behaviour are intrinsic to the broader cognitive profile of bvFTD.

Our application of a machine learning classifier was a first exploration in using FLAT variables for categorization. While the overall accuracy is encouraging, more work is needed to understand its errors and to build better models. Some errors may reflect the nature of the dementia subtypes. Of the PPAs, the model was most accurate for svPPA and nfvPPA, where language changes are more observable in spontaneous language production, and less accurate for lvPPA, where sentence repetition difficulties are an important diagnostic feature (Marshall et al., 2018). The classifier miscategorized 44% of the lvPPA sample as another type of PPA, which is consistent with the overlaps observed in group comparisons and clinicians' difficulties in assessing the clinical presentation of lvPPA (Marshall et al., 2018).

The model had the most difficulties with bvFTD, suggesting that within our study, this non-PPA subtype has the least defined language production profile. However, the bvFTD group also had the smallest sample size, and the relatively small number of data points that were available to training the algorithm may have affected accuracy. Finally, a large number of NCs were classified under svPPA. Future studies will benefit from expanding sample size and including more biographical variables in order to identify further factors that affect language in older adults. Only very large databases will be able to explore the effects of education, socio-economic status and gender in conjunction with age.

Future work can also experiment with different language elicitation tasks, more language variables and more dementia groups. To date, there is no full integrative language model for dementias which includes properties of the acoustic signal, such as prosody and speech pauses (Angelopoulou et al., 2018; López-de-Ipiña et al., 2015; Nevler et al., 2017), as well as phonological, lexical and grammatical properties. Furthermore, frequency-based variables have not been compared with generativist concepts, such as canonicity and locality, in order to reach an optimal model (as well as to test the validity of both frameworks).

Because of hitherto under-researched biographical factors it may be that, ultimately, we should not look at absolute values at a specific timepoint, but the trajectory of change within an individual. The most effective use of language as a marker of cognitive function may involve longitudinal data in conjunction with a multifactorial model of language change in ageing. Measures of information gain will be able to help select which combination of variables and standardized tests is of most use.

For clinical purposes, a practical model will contain variables that can be automatically extracted in order to minimize workload and costs. Computerized analysis can be fast and less susceptible to bias, as no raters are involved. Recordings of a few hundred words, as used in this study, can be made quickly and at bedside. As transcription technology advances, these methods will become increasingly practical and widely available for clinical applications. This in turn would open up the potential use of linguistic behavior as a biomarker to detect more subtle and ecologically relevant dysfunction of the cognitive systems, and to track this over time.

List of abbreviations

BNC: British National Corpus

BPVS: British Picture Vocabulary Scale

bvFTD: behavioral variant frontotemporal dementia

FLAT: Frequency in Language Analysis Tool

GDA: Graded Difficulty Arithmetic

GNT: Graded Naming Test

lvPPA: logopenic variant primary progressive aphasia

MMSE: Mini Mental State Examination

NCs: normal controls

nfvPPA: non-fluent variant primary progressive aphasia

PPA: primary progressive aphasia

RMT: Recognition Memory Test

SVM: support vector machine

svPPA: semantic variant primary progressive aphasia

VOSP: Visual Object Space Perception

WASI: Wechsler Abbreviated Scale of Intelligence

Consent for publication: Not applicable.

Availability of data and materials: Original data and test materials will be shared in secure, anonymised form (to protect patient confidentiality) with researchers based at other academic institutions, following an email request to the corresponding author. This may include a material transfer agreement to cover data exchange between the relevant institutions.

Competing interests: The authors declare that they have no competing interests.

Funding: This study was funded by a McKesson UK and Alzheimer's Society's grant (grant AS-PG-16-004 [Rosemary Varley and Vitor Zimmerer]). The Dementia Research Centre is supported by Alzheimer's Research UK, Brain Research Trust, and The Wolfson Foundation. Individual authors were supported by an Action on Hearing Loss–Dunhill Medical Trust Pauline Ashley Fellowship (grant PA23_Hardy [Chris Hardy]), a Medical Research Council PhD Studentship (Rebecca Bond), a Medical Research Council Clinician Scientist Fellowship (Jonathan Rohrer). Jason Warren is supported by an Alzheimer's Society (grant AS-PG-16-007) and the NIHR UCL/UCLH Biomedical Research Centre.

Acknowledgments: We thank Michael Coleman for programming the FLAT.

References

Angelopoulou, G., Kasselimis, D., Makrydakis, G., Varkanitsa, M., Roussos, P., Goutsos, D., … Potagas, C. (2018). Silent pauses in aphasia. *Neuropsychologia*, *114*, 41–49. https://doi.org/10.1016/j.neuropsychologia.2018.04.006

Ash, S., Evans, E., O'Shea, J., Powers, J., Boller, A., Weinberg, D., … Grossman, M. (2013). Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology*, *81*(4), 329–336. https://doi.org/10.1212/WNL.0b013e31829c5d0e

Avrutin, S. (2000). Comprehension of Discourse-Linked and Non-Discourse-Linked Questions by Children and Broca's Aphasics. *Language and the Brain*, 295–313. https://doi.org/10.1016/B978-012304260-6/50017-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017, March 6). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*. Frontiers Research Foundation. https://doi.org/10.3389/fpsyg.2017.00269

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207–1215. https://doi.org/10.1016/S0028-3932(00)00034-8

Bridges, K. A., & Van Lancker Sidtis, D. (2013). Formulaic language in Alzheimer's disease. *Aphasiology*, *27*(7), 799–810. https://doi.org/10.1080/02687038.2012.757760

Bruns, C., Varley, R. A., Zimmerer, V. C., Carragher, M., Brekelmans, G., & Beeke, S. (2019). "I don't know": a usage-based approach to familiar collocations in non-fluent aphasia. *Aphasiology*, *33*(2), 140–162. https://doi.org/10.1080/02687038.2018.1535692

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Code, C. (1982). Neurolinguistic analysis of recurrent utterance in aphasia. *Cortex*, *18*, 141–152. https://doi.org/10.1016/S0010-9452(82)80025-7

Code, C. (2013). Letter to the editor and author's response: Did Leborgne have one or two speech automatisms? *Journal of the History of the Neurosciences*, *22*, 319–320.

Conklin, K., & Schmitt, N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale (BPVS-II) 2nd ed. NFER-Nelson Publishing Company*.

Folstein, M., Folstein, S., & McHugh, P. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *3*, 189–198.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, *55*, 43–60. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010945212003413

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease : JAD*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Gahl, S, Menn, L., Ramsberger, G., Juracsky, D., Elder, E., Rewega, M., & Audrey, L. H. (2003). Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences. *Brain and Cognition*, *53*, 223–228. https://doi.org/https://doi.org/10.1016/S0278-2626(03)00114-3

Gahl, Susanne, & Menn, L. (2016). Usage-based approaches to aphasia. *Aphasiology*, 1–17. https://doi.org/10.1080/02687038.2016.1140120

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *TRENDS in Cognitive Sciences*, *7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., … Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, *76*(11), 1006–1014. https://doi.org/10.1212/WNL.0b013e31821103e6

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269–291). Frankfurt am Main: Peter Lang.

Grodzinsky, Y. (2000). The Neurology of Syntax: Language Use Without Broca's Area. *Behavioral and Brain Sciences*, *23*, 1–71. https://doi.org/10.1017/S0140525X00002399

Hardy, C. J. D., Buckley, A. H., Downey, L. E., Lehmann, M., Zimmerer, V. C., Varley, R. A., … Warren, J. D. (2015). The language profile of behavioral variant frontotemporal dementia. *Journal of Alzheimer's Disease*, *50*(2). https://doi.org/10.3233/JAD-150806

Hatchard, R., & Lieven, E. (2019). Inflection of nouns for grammatical number in spoken narratives by people with aphasia: A challenge for rule-based models. *Language and Cognition*, *11*(3), 341–372. https://doi.org/https://doi.org/10.1017/langcog.2019.21

Huppert, F. A., Cabelli, S. T., Matthews, F. E., & MRC Cognitive Function and Ageing Study, the M. C. F. and A. S. (MRC. (2005). Brief cognitive assessment in a UK population sample -- distributional properties and the relationship between the MMSE and an extended mental state examination. *BMC Geriatrics*, *5*, 7. https://doi.org/10.1186/1471-2318-5-7

Jackson, M., & Warrington, E. K. (1986). Arithmetic Skills in Patients with Unilateral Cerebral Lesions. *Cortex*. https://doi.org/10.1016/S0010-9452(86)80020-X

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features (pp. 137–142). Springer, Berlin, Heidelberg. https://doi.org/10.1007/bfb0026683

Keady, J., & Gillard, J. (2002). The experience of neuropsychological assessment for people with suspected Alzheimer's disease. In *The person with Alzheimer's disease: Pathways to understanding the experience* (pp. 3–28).

López-de-Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., … Eguiraun, H. (2015). On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature. *Cognitive Computation*, *7*(1), 44–55. https://doi.org/10.1007/s12559-013-9229-9

Marshall, C. R., Hardy, C. J. D., Volkmer, A., Russell, L. L., Bond, R. L., Fletcher, P. D., … Warren, J. D. (2018). Primary progressive aphasia: a clinical approach. *Journal of Neurology*, *265*(6), 1474–1490. https://doi.org/10.1007/s00415-018-8762-6

MATLAB and Statistics and Machine Learning Toolbox. (2018). Natick, Massachusetts: The MathWorks, Inc.

Mauner, G., Fromkin, V. A., & Cornell, T. L. (1993). Comprehension and acceptability judgments in agrammatism: Disruptuions in the syntax of referencial dependency. *Brain and Language*, *45*, 340–370.

McKenna, P., & Warrington, E. K. (1983). *Graded Naming Test*. Windsor, UK: NFER-Nelson.

Menn, L., Gahl, S., Holland, L., Ramsberger, G., & Jurafsky, D. S. (2003). Beyond canonical form: Verb-frame frequency affects verb production and comprehension. *Brain and Language*, *87*, 23–24.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*(1–2), 169–186. https://doi.org/10.1016/S0925-2312(03)00431-4

Mondini, S., Jarema, G., Luzzatti, C., Burani, C., & Semenza, C. (2002). Why is “Red Cross” different from “Yellow Cross”?: A neuropsychological study of noun-adjective agreement within Italian compounds. *Brain and Language*, *81*(1–3), 621–634. https://doi.org/10.1006/brln.2001.2552

Nevler, N., Ash, S., Jester, C., Irwin, D. J., Liberman, M., & Grossman, M. (2017). Automatic measurement of prosody in behavioral variant FTD. *Neurology*, *89*(7), 650–656. https://doi.org/10.1212/WNL.0000000000004236

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Rascovsky, K., Hodges, J., Knopman, D., Mendez, M., Kramer, J., Neuhaus, J., … Miller, B. (2011, September 1). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *BRAIN*. OXFORD UNIV PRESS. Retrieved from http://discovery.ucl.ac.uk/1325627/

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. B. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language*, *175*, 111–122. https://doi.org/10.1016/j.bandl.2017.10.004

The British National Corpus, version 2 (BNC XML Edition). (2007). Retrieved from http://www.natcorp.ox.ac.uk

Van Lancker Sidtis, D. (2012). Formulaic Language and Language Disorders. *Annual Review of Applied Linguistics*, *32*, 62–80. https://doi.org/10.1017/S0267190512000104

Van Lancker Sidtis, D., & Yang, S. (2016). Formulaic language performance in left- and right-hemisphere damaged patients: structured testing. *Aphasiology*, *31*(March), 1–18. https://doi.org/10.1080/02687038.2016.1157136

Warrington, E. K. (1984). *Recognition Memory Test: Manual*. Berkshire, UK: NFER-Nelson.

Warrington, E. K., & James, M. (1991). *The Visual Object and Space Perception Battery*. *Thames Valley Test Company*. https://doi.org/Thesis_references-Converted #318

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence (WASI-II)*. San Antonio, TX: NCS Pearson.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., & Woo, K. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Soringer-Verlag.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., … Gorno-Tempini, M. L. (2010). Connected speech production in three variants of temporary progressive aphasia. *Brain*, *133*, 2069–2088.

Wray, A. (2011). Formulaic language as a barrier to effective communication with people with Alzheimer’s Disease. *The Canadian Modern Language Review*, *67*(4), 429–458. https://doi.org/10.3138/cmlr.67.4.429

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, *32*, 231–254. https://doi.org/10.1017/S026719051200013X

Wray, A. (2014). Dementia and language. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*.

Oxford: Wiley-Blackwell.

Zimmerer, V. C., Newman, L., Thomson, R., Coleman, M. J., & Varley, R. A. (2018). Automated analysis of language production in aphasia and right hemisphere damage: Frequency and collocation strength. *Aphasiology*, *32*(11), 1267–1283. https://doi.org/10.1080/02687038.2018.1497138

Zimmerer, Vitor C., Dąbrowska, E., & Varley, R. A. (n.d.). The syntax-lexicon continuum: Explaining variation in aphasic language. *Journal of Cognitive Linguistics*.

Zimmerer, Vitor C, Wibrow, M., & Varley, R. A. (2016). Formulaic language in people with probable Alzheimer's Disease: a frequency-based approach. *Journal of Alzheimer's Disease*, *53*, 1145–1160. https://doi.org/10.3233/JAD-160099

Figures

Figure 1. Radar plot visualization of language profiles of different dementia groups.

Data were residualized over participant age in order to account for age differences, and then normalized using control means and standard deviations. The outer line in each plot represents the control mean; each line towards the center represents a distance of one standard deviation from the control mean.

Figure 2. Correlation heatmaps displaying the relationship between spontaneous language output and standardized measures of language and cognition.

Colors indicate effect size (Pearson’s r). Tests were classified according to whether administration and stimuli are predominantly verbal or non-verbal. The MMSE was categorized as a test of general cognition, and was not administered to controls.

Figure 3. Classification matrix based on the SVM classifier.

Columns display the clinical diagnosis. Rows show the predicted class on the basis of the five FLAT variables. Percentages represent the proportion of members of the true class with a given diagnosis. Beneath is a raw count of the same members.

Appendices

Appendix A. Report of significant group differences.

There was a significant difference in MMSE scores between patient groups, F(1,75) = 4.77, p = .03; people with svPPA had significantly lower scores than the nfvPPA group, (p = .012) and the lvPPA group lower scores than the nfvPPA group (p = .007).

There were differences between groups in all subtests of the Wechsler Abbreviated Scale of Intelligence (WASI). WASI vocabulary scores differed significantly between groups, F(1,99) = 52.80, p < .001, as all patient groups had significantly lower scores than controls (p < .01). WASI block scores were significantly different between groups, F(1,99) = 5.61, p = .019; all patient groups had significantly lower scores than NCs (p < .001). WASI Matrix scores differed significantly between groups, F(1,99) = 6.38, p = .01; all patient groups had significantly lower scores than NCs (p < .001). Scores for WASI Similarities differed significantly across groups, F(1,99) = 44.85, p < .001; all patient groups were significantly worse than NCs (p < .001).

All patient groups performed worse on the Recognition Memory Test (RMT),(Warrington, 1984) which assesses episodic memory. There were significant differences in RMT Words scores between groups, F(1,86) = 35.67, p < .001; all patient groups were significantly worse than NCs (p < .05). RMT Faces scores also differed significantly between groups, F(1,92) = 12.48, p < .001; all patient groups performed significantly worse than NCs (p < .05). Maximum forward digit span (Wechsler, 2011), which is an indicator of verbal working memory, also differed significantly between groups, F(1,96) = 4.99, p = .028; lvPPA and nfvPPA groups performed significantly worse than controls (p < .001). There was no main effect for backwards digit span as an indicator of executive function (Wechsler, 2011), F(1,94) = 1.87, p = .175, however, pairwise comparisons showed significant differences between NCs and lvPPA, nfvPPA and bvFTD groups (p < .05).

Groups differed in their capacity for calculation, which was assessed using the Graded Difficulty Arithmetic (GDA),(Jackson & Warrington, 1986) F(1,89) = 15.37, p < .001; all patient groups performed worse than NCs (p < .05). Object and space perception was measured using the Visual Object Space Perception (VOSP) (Warrington & James, 1991). There was no main group effect for VOSP scores, F(1,98) = 2.24, p = 0.137; however, pairwise comparisons showed that the lvPPA group performed worse than NCs (p < .05).

There were significant group differences for both naming and word comprehension: Groups differed significantly the Graded Naming Test (GNT), (McKenna & Warrington, 1983) F(1,97) = 145.16, p < .001, as all groups performed more poorly than NCs (p < .001). Performance in the British Picture Vocabulary Scale (BPVS) (Dunn et al., 1997) also differed between groups, F(1,96) = 31.72, p < .001; svPPA and lvPPA groups had lower scores than NCs (p < .05).

Appendix B. Pairwise comparisons of properties of spontaneous language output, with age entered as covariate. Suggested Bonferroni adjustment of significance threshold due to ten comparisons for each variable/hypothesis: p < .005. MD = mean difference, calculated as: mean of group in the left column subtracted by the mean of the group in the top row. PPA = primary progressive aphasia; FTD = fronto-temporal dementia.

Appendix B.1 Word count.

Word count	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioral variant FTD

Normal controls	MD = 12
	95% CI [-60, 83]
	p = .749
		MD = 106
		95% CI [32, 180]
		p = .006
			MD = 177
			95% CI [105, 248]
			p < .001
				MD = 145
				95% CI [59, 231]
				p = .001

Semantic variant PPA		MD = 94.124
		95% CI [27, 162]
		p = .007
			MD = 165
			95% CI [101, 229]
			p < .001
				MD = 133
				95% CI [53, 214]
				p = 001

Logopenic variant PPA			MD = 71
			95% CI [4, 138]
			p = .039
				MD = 40
				95% CI [-43, 122]
				p = .346

Non-fluent variant PPA				MD = -31
				95% CI [-111, 48]
				p = .436

Appendix B.2 Combination ratio.

Combination ratio	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioral variant FTD

Normal controls	MD = .08
	95% CI [.01, .14]
	p = .021
		MD = .18
		95% CI [.11, .24]
		p < .001
			MD = .35
			95% CI [.29, .41]
			p < .001
				MD = .18
				95% CI [.1, .25]
				p < .001

Semantic variant PPA		MD = .1
		95% CI [.04, .16]
		p = .001
			MD = .27
			95% CI [.22, .33]
			p < .001
				MD = .1
				95% CI [.03, .17]
				p = 006

Logopenic variant PPA			MD = .17
			95% CI [.11, .23]
			p < .001
				MD = -.001
				95% CI [-.07, .07]
				p = .985

Non-fluent variant PPA				MD = -.17 95% CI [-.24, -.1] p < .001

Appendix B.3 Content word ratio.

Content word ratio	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioral variant FTD

Normal controls	MD = .38
	95% CI [.002, .075]
	p = .039
		MD = .09
		95% CI [.05, .13]
		p < .001
			MD = .03
			95% CI [-.002, .07]
			p = .065
				MD = .05
				95% CI [.003, .09]
				p = .035

Semantic variant PPA		MD = .05
		95% CI [.02, .09]
		p = .004
			MD = -.004
			95% CI [-.04, .03]
			p = .8
				MD = .009
				95% CI [-.03, .5]
				p = .67

Logopenic variant PPA			MD = -.06
			95% CI [-.09, -.02]
			p = .002
				MD = -.04
				95% CI [-.08, 0]
				p = .048

Non-fluent variant PPA				MD = .01
				95% CI [-.03, .05]
				p = .525

Appendix B.4 Content word frequency.

Content word frequency	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioral variant FTD
Normal controls	β = 6985, SE = 1579, $\chi^2(4)$ = 16.992, p < .001	β = 11063, SE = 1621, $\chi^2(4)$ = 24.303, p < .001	β = 7845, SE = 1279, $\chi^2(4)$ = 22.726, p < .001	β = 1753, SE = 468, $\chi^2(4)$ = 22.6, p < .001
Semantic variant PPA		β = 14897, SE = 3768, $\chi^2(4)$ = 15.581, p < .001	β = 8392, SE = 2148, $\chi^2(4)$ = 15.216, p < .001	β = -161, SE = 1131, $\chi^2(4)$ = .02, p = .89
Logopenic variant PPA			β = 3737, SE = 6806, $\chi^2(4)$ = .301, p = .58	β = -8828, SE = 3385, $\chi^2(4)$ = 7.19, p = .007
Non-fluent variant PPA				β = -19762, SE = 7667, $\chi^2(4)$ = 6.623, p = .01

Appendix B.5 Bigram collocation strength.

Bigram collocation strength	Semantic variant PPA	Logopenic variant PPA	Non-fluent variant PPA	Behavioral variant FTD
Normal controls	β = 4.77, SE = .99, $\chi^2(4) = 17.882$, $p < .001$	β = 4.92, SE = .85, $\chi^2(4) = 24.448$, $p < .001$	β = 2.99, SE = .56, $\chi^2(4) = 18.892$, $p < .001$	β = 1.04, SE = .38, $\chi^2(4) = 6.62$, $p = .01$
Semantic variant PPA		β = 4.69, SE = 1.43, $\chi^2(4) = 9.871$, $p = .002$	β = 2.03, SE = .67, $\chi^2(4) = 8.223$, $p = .004$	β = -.04, SE = .46, $\chi^2(4)=.01$, $p=.94$
Logopenic variant PPA			β = -.35, SE = 2.16, $\chi^2(4) = .02$, $p = .87$	β = -3.05, SE = 1.2, $\chi^2(4) = 5.847$, $p = .02$
Non-fluent variant PPA				β = -5.41, SE = 2.19, $\chi^2(4) = 5.673$, $p = .017$

Appendix C. Correlations (Pearson's r, two-tailed) at p < .05 between FLAT measures and established measures of dementia progression.

Appendix C.1 Normal controls. Number of data points: All measures (20).

FLAT variable	**Correlations with verbal measures**	**Correlations with non-verbal measures**
Word count	-	-
Combination ratio	-	-
Content word ratio	-	-
Content word frequency	-	RTM Faces, r = .569, p = .009
Collocation strength (bigram t-score)	-	RTM Faces, r = .472, p = .036

Appendix C.2 Semantic variant PPA: Number of data points: WASI Vocab (29), WASI Blocks (25), WASI Similarities (25), WASI Matrices (25), RMT Faces (22), RMT Words (21), BPVS (29), GNT (24), VOSP (25), Digit Span forward (25), Digit Span backward (25), Arithmetic total (24).

FLAT variable	**Correlations with verbal measures**	**Correlations with non-verbal measures**
Word count	Digit Span backward, r = .436, p = .029	-
Combination ratio	WASI Similarities, r = .426, p = .034; Digit Span forward, r = .409, p = .043; Digit Span backward, r = .589, p = .002	Arithmetic, r = .577, p = .003
Content word ratio	-	WASI Matrices, r = .424, p = .035

Content word frequency	WASI Vocab, r = -.45, p = .014; WASI Similarities, r =-.429, p = .032, RMT Words, r = -.571, p = .007; Digit Span backwards, r = -.424, p = .035	WASI Blocks, r = -.583, p = .002; WASI Matrices, r = -.666, p < .001; VOSP, r = -.428, p = .033; Arithmetic, r = -.546, p = .006
Collocation strength (bigram t-score)	-	-

Appendix C.3 Logopenic variant PPA: Number of data points: WASI Vocab (20), WASI Blocks (20), WASI Similarities (20), WASI Matrices (20), RMT Faces (19), RMT Words (16), BPVS (19), GNT (20), VOSP (20), Digit Span forward (20), Digit Span backward (20), Arithmetic total (17).

FLAT variable	**Correlations with verbal measures**	**Correlations with non-verbal measures**
Word count	GNT, r = .459, p = .042	-
Combination ratio	-	-
Content word ratio	Digit Span forward, r = .621, p = .003	-
Content word frequency	Digit Span backward, r = -.526, p = .017	WASI Matrices, r = -.468, p = .037; RMT Faces, r = -.457, p = .049
Collocation strength (bigram t-score)	Digit Span forward, r = -.576, p = .008	-

Appendix C.4 Non-fluent variant PPA: Number of data points: WASI Vocab (22), WASI Blocks (22), WASI Similarities (22), WASI Matrices (22), RMT Faces (20), RMT Words (19), BPVS (21), GNT (21), VOSP (21), Digit Span forward (20), Digit Span backward (18), Arithmetic total (18).

FLAT variable	**Correlations with verbal measures**	**Correlations with non-verbal measures**
Word count	-	-
Combination ratio	-	-
Content word ratio	WASI Vocabulary, r = .465, p = .029; WASI Similarities, r = -.496, p = .019	-
Content word frequency	WASI Vocabulary, r = -.703, p < .001; BPVS, r = -.517, p = .016; GNT, r = -.726, p < .001; Digit Span forward, r = -.52, p = .019	-
Collocation strength (bigram t-score)	WASI Vocabulary, r = -.579; GNT, r = -.519, p = .016; Digit Span forward, r = .666, p = .001; Digit Span backward, r = .532, p = .023	-

Appendix C.5 Behavioral variant FTD: Number of data points: WASI Vocab (14), WASI Blocks (14), WASI Similarities (14), WASI Matrices (14), RMT Faces (13), RMT Words (12), BPVS (14), GNT (14), VOSP (14), Digit Span forward (13), Digit Span backward (13), Arithmetic total (12).

FLAT variable	**Correlations with verbal measures**	**Correlations with non-verbal measures**
Word count	-	RMT Faces, r = .659, p = .014
Combination ratio	-	-
Content word ratio	WASI Vocabulary, r = .653, p = .011; WASI Similarities, r = .627, p = .016; Digit Span backward, r = .719, r = .006	-
Content word frequency	WASI Similarities, r = -.603, p = .022	-
Collocation strength (bigram t-score)	-	-