Automated Analysis of Language Production in Aphasia and Right Hemisphere Damage: Frequency and Collocation Strength

Zimmerer, V. C.[1], Newman, L.[2], Thomson, R.[3], Coleman, M.[1], Varley, R. A.[1]

[1]Department of Language and Cognition, University College London, London, UK

[2]Department of Clinical, Educational and Health Psychology, University College London, London, UK

[3]Royal Free Hospital, Royal Free London NHS Foundation Trust, London, UK

Corresponding author:

Vitor Zimmerer

Chandler House

2 Wakefield Street

London WC1N 1PF

E-mail: v.zimmerer@ucl.ac.uk

Abstract

Background: Reliance on formulaic language, i.e. holistically processed multiword chunks, is claimed to distinguish speakers with aphasia, speakers with right hemisphere damage and neurotypical controls. Frequency and collocation strength of word combinations are indicators of formulaic language.

Methods and Procedures: We used computerized methods to investigate spontaneous language in 40 speakers: 10 with fluent aphasia, 10 with non-fluent aphasia, 10 with right-hemisphere damage and 10 neurotypical controls. Our analysis focused on frequency and collocation strength of grammatical combinations as markers of formulaic language (using the British National Corpus as reference), but also looked at word frequency, language fluency, proportion of content words, and measures of lexical and combinatorial diversity.

Results: Both groups with aphasia differed from neurotypical speakers with regard to lexical features and word combinations. Their language was less fluent, less diverse at word level, and in the non-fluent group, contained a higher proportion of content words. Each aphasic group also differed from controls with increased values on at least one marker of formulaic language. Speakers with right hemisphere damage produced less frequent combinations which were more weakly collocated; however, these effects did not reach statistical significance.

Conclusions: Our results show that formulaic language use distinguishes aphasic from other speakers, and that differences can be tracked with an automated, corpus-based script which uses frequency and collocation variables. We present our study in the context of usage-based frameworks of language.

Keywords: Aphasia; Right hemisphere damage; Automated language analysis; Frequency; Collocation strength

1. Introduction

It is well established that words which a speaker frequently encounters are processed more quickly and accurately (Dell, 1990; Harley & Bown, 1998; Jescheniak & Levelt, 1994). Frequency has become an important variable with which to capture the individual's language experience, and interacts with a range of other factors including age of acquisition, valence, word length and semantic diversity and neighbourhood density (Baayen, Milin, & Ramscar, 2016). Frequency effects extend beyond the single word level to combinatorial frequencies of multiword expressions. Frequency affects processing speed and accuracy of word combinations even if they have the same formal syntactic structure (Arnon & Snider, 2010; Conklin & Schmitt, 2008; Jacobs, Dell, & Bannard, 2017; Janssen & Barber, 2012; Tremblay & Baayen, 2010). For instance, the utterance *I like it* is easier to produce than *I keep it*. These effects can be seen as one facet of the phenomenon known as formulaic language (Code, 2005; Conklin & Schmitt, 2012; Wray, 2002, 2012; Wray & Perkins, 2000). Language formulas are familiar multiword units that operate holistically either via strengthened connections between individual words or through representation as single lexical items. This sets formulas apart from novel or non-idiomatic utterances which may rely more on combinatorial and analytical processes. An analysis of event-related potentials (Siyanova-Chanturia, Conklin, Caffarra, Kaan, & van Heuven, 2017) supports the view that formulas are processed differently: they examined reading of familiar NOUN *and* NOUN combinations (e.g., *knife and fork*) and less familiar combinations (e.g., *spoon and fork*). Familiar combinations elicited a stronger P300, associated with context expectancy and phrasal processing, and a smaller N400, interpreted as easier semantic integration. Effects disappeared when noun combinations were presented without the conjunction (e.g., *knife - fork*), suggesting that they depend on the particular word combinations and not merely on the co-occurrence of content words.

While every speaker uses formulaic language, the degree of formulaicity can change as the result of neurological damage. Disruption of lexical-semantic and/or syntactic networks is associated with reduced production of novel, creative language, and an increased tendency to use formulaic language. In Code's (1982, 1983, 1989) collection and analysis of recurrent utterances in aphasia, many meaningful combinations were formulas such as *wait a minute*, *pen and paper*, or *cor blimey*. Blanken (1991) made similar observations, and noted that high familiarity and frequency were markers of these expressions. In comparison with controls, studies have found increased formulaicity in the language output of speakers with aphasia and Alzheimer's disease (Bridges & Van Lancker Sidtis, 2013; Van Lancker Sidtis & Postman, 2006; Zimmerer, Wibrow, & Varley, 2016). Jaecks, Hielscher-Fastabend and Stenneken (2012) also found increased use of formulaic language as a marker of residual aphasia. Results demonstrate that formulaic language is not only easier to process for healthy speakers, but also resilient to some types of damage. Formulas can therefore be seen as enabling communication, however they also limit the scope of communication and are less suited for novel situations or expressing new ideas (Wray, 2011). While an increase in formulaic language has been observed in aphasia and dementia, there are reports of decreased use of formulas in speakers with right hemisphere damage (Van Lancker Sidtis & Postman, 2006) and also in Parkinson's disease (Van Lancker Sidtis, Choi, Alken, & Sidtis, 2015). Van Lancker and collaborators interpret these data as support for a dual-process model, in which novel/analytic language is largely dependent upon left hemisphere computational mechanisms, while formulas are represented in the right hemisphere and subcortical structures (see also Van Lancker Sidtis, 2012).

These studies indicate that in addition to established markers such as agrammatism, paragrammatism, paraphasia and changes in structural complexity, the degree to which an individual relies on formulaic language is a relevant aspect of profiling impaired language.

However, there have only been few investigations in individuals with brain damage. We report new findings from people with fluent aphasia (*fA*) and non-fluent aphasia (*nfA*), as well as non-aphasic people with right-hemisphere damage (*RHD*) and neurotypical controls (*NC*). In the studies by Van Lancker and colleagues (see above) analysis was carried out by hand through marking utterances considered formulaic. The degree of formulaicity was determined by calculating the proportion of words in a language sample flagged as part of a formula. Our methods are substantially different. We developed a computer program, the Frequency in Language Analysis Tool (*FLAT*) which extracts every word, bigram (two-word-combination) and trigram (three-word-combination) from an orthographic text, e.g. a transcript, and determines the usage frequency of these units. As a reference it uses the spoken subcorpus of the British National Corpus (*BNC*, 2007) which consists of 10 million words from different talk contexts, geographic regions and demographic groups. The assumption is that frequency and collocation patterns in the corpus reflect everyday use. Based on these values, the FLAT can be used to explore the extent of use of common or strongly collocated combinations in an individual or group. The method is different from many corpus linguistic approaches where usage properties are retrieved for select words or combinations, for example for contrasting high/low frequency conditions in a psycholinguistic experiment. Instead, the FLAT can extract values for every unit produced in a sample in order to estimate an overall degree of formulaicity. This procedure, as well as the use of a reference corpus to not only extract frequencies of words, but also of larger units, distinguishes FLAT analysis from a program such as AntConc (Anthony, 2018).

Our approach regards FLAT extracted frequency and collocation strength values as indicators of formulaicity in individuals or groups. Because the FLAT cannot distinguish grammatical from ungrammatical combinations, we exclude the latter at the annotation stage (see Methods). Our work with FLAT therefore addresses the question of whether in cases of grammatical combinations, individuals or groups rely more on common and more likely to be formulaic, or less common and more likely to be analytically processed, forms. Using this approach, Zimmerer, Wibrow and Varley (2016) used the FLAT to investigate Cookie Theft picture descriptions produced by people with probable Alzheimer's disease and neurotypical controls. Samples were taken from DementiaBank's Pitt corpus (Becker, Boller, Lopez, Saxton, & McGonigle, 1994; MacWhinney, 2007). For grammatical combinations, average frequency and t-score (a measure of collocation strength) were higher in the Alzheimer's group. Further, speakers who were longer post-diagnosis produced more frequent and strongly collocated combinations. Frequency effects were much stronger at the level of word combinations than they were at the single word level, albeit the analysis did not include a comparison for content words alone, which are usually the focus of studies on lexical processing (Cuetos, Arce, Martínez, & Ellis, 2015; Cuetos, Rodríguez-Ferreiro, Sage, & Ellis, 2012).

We see a number of advantages in using the FLAT programme for profiling the degree of formulaic language. The program itself is blind to hypotheses and group membership, which reduces the risk of bias. Further, its outputs are continuous variables instead of binary decisions on whether a given expression is considered formulaic. The FLAT is fast, being able to process more than 100 words per second, although samples need to be formatted for FLAT which requires some informed choices (see Methods and Supplement). The tool has some limitations: It is blind to semantic and pragmatic properties of combinations. While it is therefore well-suited for capturing strongly collocated or high-frequency lexical bundles, the method is not reliable for detecting idioms and it is not possible to determine formula types on the basis of its output. FLAT also treats each specific word combination as unique and does not group variations, e.g. in inflection (e.g., *drive* vs. *drives*) or contraction (e.g., *don't* vs. *do not*). Finally, and as noted above, it cannot detect ungrammatical combinations, and if included these will be analysed together with grammatical combinations, making it impossible to determine whether low values are the result of rare or creative, grammatical combinations, or errors. With these properties in mind our project not only aimed to profile language in people with brain damage, but also to test the ability of FLAT to distinguish patient groups.

In this study we measured average frequency and collocation strength of combinations for each individual, and also average frequency of content words. We also assumed that fluent and non-fluent aphasic speakers would differ from one another with regards to formulaicity, albeit with no directional hypothesis. Previous studies on formulaic language in aphasia have not tested this distinction. In addition, we determined the variation within words, bigrams and trigrams using type-token-ratios (*TTR*). While TTR is typically used to analyse lexical diversity, Perkins (1994) also used it at a multiword level to analyse diversity in combinations. Our hypotheses for aphasic groups were based on observations of increased formulaicity in aphasic samples. We hypothesized that frequencies for words and grammatical word combinations would be higher than in controls. We also expected that variety of expressions, fluency (defined as ratio of words appearing in connected units) and number of words would be reduced for both aphasic groups. We expected people with nfA to produce proportionally more content words than the other groups.

We further hypothesized that word combinations would have lower frequency and collocation strength in the RHD group than in the control group, and consequently be less formulaic than speakers with aphasia. With regards to diversity of words and word combinations, verbal fluency and word count, we hypothesized a difference between RHD and NC groups (in any direction), and higher values for speakers with RHD than speakers with aphasia. However, due to the small number of studies, the evidence base for hypotheses regarding speakers with RHD damage is weaker than for aphasic speakers.

2. Methods

2.1 Participants

Samples were originally collected in Sheffield, South Yorkshire, UK, for an analysis of deictic forms by Varley (1993). Participants gave written consent to participation in the research and to the long-term storage of their anonymized data. The original corpus included 20 NC speakers, 20 non-aphasic individuals with RHD, 20 speakers with fA and 14 with non-fluent aphasia nfA. Samples from 20 aphasic participants (10 from each aphasic sub-group) were available in machine readable form from a previous study. A further 10 samples were selected via a random number generator from each of the non-aphasic groups for conversion into electronic format and further analysis. Biographical characteristics of the final groups were: fA, mean age = 57.8, SD = 9.9; 5 male; nfA, 51.3, SD = 9.6; 6 male; RHD, mean age = 59.4, SD = 14.38; 5 male; NC; mean age = 72.4, SD = 7.53; 5 male. Diagnosis of fluent/non-fluent aphasia was made by the referring speech and language therapist. Criteria focused on linguistic features rather than phonetic speech fluency: non-fluent participants had grammatical difficulties, while fluent aphasic participants had predominant lexical-semantic impairments. All lesions were of vascular origin, and the lesioned hemisphere was determined on the basis of clinical signs, such as hemiplegia. All participants with aphasia were diagnosed to have left-hemisphere lesions, with exception of one person with nfA who had damage to the right hemisphere. No participant with RHD was diagnosed with aphasia. Participant groups differed by age, $F(3,36) = 6.891$, $p > .001$, as NCs were significantly older than fA ($p = .004$), nfA ($p < .001$) and RHD groups ($p = .01$).

2.2 Procedure

Varley conducted a semi-structured interview with each participant. Topics were history of illness, family, occupational and life history. Each sample was ca. 30 minutes long and recorded on audio tape. For the current study, interview transcripts had to be copied from hand writing or typed transcripts into text files, and in some cases audio had to be transcribed anew. New transcriptions were made using F4transkript (2012).

The FLAT parser requires additional formatting of transcripts. As a first step, we removed non-words such as neologisms, as the BNC has no information about them. When we judged a non-word as a phonological error and were able to trace a clear relationship with an intended target, we changed the word to that target so that FLAT could recognize it. We also removed ungrammatical word combinations. Because we interpret low frequency units as rare or creative, our analysis would be confounded if non-words or combinatorial errors were included. Such errors would be assigned low frequency values (e.g., *he walk*) and would therefore be classified as highly creative. To avoid this, we inserted "separator" tags into ungrammatical combinations (*he <.> walk*) which the FLAT interprets as markers for exclusion from bigram analysis. In cases of false starts we also added separators (e.g., *did you <.> were you involved?*). Note that with this method allows investigation of parts of production which are grammatical, i.e. it is not necessary to exclude entire utterances. Sentence final punctuation also served as exclusion markers as bigrams which started at the end of one sentence and ended at the beginning of the next sentence were not analysed. In the case of immediate repetitions, we deleted the first instance (e.g., *she she went* -> *she went*), unless they were judged to be intentional (e.g., *very, very wide*). As a result, FLAT analysis concerns usage properties of real words and grammatical combinations exclusively.

Annotated transcripts were then run through FLAT version 2 (Zimmerer, Coleman, Wibrow, 2017) which extracts raw frequencies for words, bigrams and trigrams, extracting combinations by moving bi- and trigram sized "windows" through the text one word at a time (Table 1). It also counts the number of tokens and types for each level. It then generates a range of variables that characterize language samples (Table 2). It first computes a "combination ratio", a variable indicating fluency or connectedness of the sample, by dividing the number of trigrams by the number of words. High values indicate that output is characterized by combinations of three or more words rather than single words or two-word combinations. It then calculates TTR for words, bigrams and trigrams. High values indicate repeated use of a small number of words or word combinations.

FLAT then averages raw frequencies of word, bigram and trigram types for each individual. Using a list of closed class words, it categorizes individual words as closed class (function) or open class (content) words and averages the frequency for content words separately. Collocation strength is measured as t-scores (Gries, 2010), which express the degree to which words in a combination are associated in language use, i.e., how often words in a combination appear together (frequency) in relation to how often they would appear if words were randomly scrambled in the corpus (expected frequency; see Table 2). *Cor blimey*, for instance, has low frequency but high collocation strength since the words appear together very often, but rarely in other combinations. In this regard t-scores are similar to Mutual Information measures, however, they are better suited for low frequency combinations (Church & Hanks, 1990; Gries, 2010).

Zimmerer, Wibrow and Varley (2016) observed that as unit size increased from words to bigrams and trigrams, so did the likelihood that a unit did not occur in the BNC, i.e., had a frequency of zero. The resulting floor effect made variables less precise. In order to reduce the number of statistical tests, Zimmerer et al. excluded trigrams from the final step of the analysis, since only 58.6% of them were in the BNC. They also calculated for each speaker the proportion of units that occur in the BNC as this represents a useful measure of productivity/creativity: A speaker who only produces combinations that occur in the reference corpus is likely to be more formulaic than a speaker whose sample includes a smaller proportion of combinations from the corpus. As in the previous study, we capture this proportion using the variable “BiBNC ratio” (proportion of bigrams which appear in the BNC). As with frequency and t-scores, higher BiBNC ratios indicate more use of formulaic combinations.

[INSERT TABLE 1 HERE]

Table 1. Frequency data extracted by FLAT from the spoken subcorpus of the BNC, demonstrated on the utterance *I don't know* (first row), which the BNC characterizes as four words (*I*, *do*, *n't*, *know*) and three bigrams (*I do*, *don't*, *n't know*). The FLAT moves through each text word-by-word and extracts raw frequencies for each level (the first value in the bigram frequency row denotes the frequency of *I do*). Numbers stand for occurrences per million words.

	I	do	n't	know
Word type	Function	function	function	content
Word frequency	309559	99559	126276	57430
Bigram frequency	-	21568	42216	10646

[INSERT TABLE 2 HERE]

Table 2. Variable list for language analysis. Variables were computed on the basis of raw values extracted using the FLAT (Table 1), on the basis of annotated transcripts.

Variable	Explanation
Word count (N of words)	Quantity of verbal output
Bigram/Trigram count (N of bigrams/trigrams)	Number of grammatical word combinations
Combination ratio	Measure of connected language. The proportion of words that occur in combinations of three words and more.
Word/Bigram frequency	Frequency of words and combinations in tokens per million words, extracted from the spoken subcorpus of the BNC. Excludes combinations with frequency = 0.
Content word frequency	Frequency of content (open class) words
Word/bigram/trigram type-token ratio (TTR)	Number of tokens divided by number of types. Seen as a measure of lexical/combination diversity.
Bigram expected frequency	Needed to calculate collocation strength. Expected frequency of a combination if

	individual word frequency was maintained, but order of words was random.
Bigram t-score	Collocation strength of combinations, taking into account individual word frequency. Excludes combinations with frequency = 0.
BiBNC ratio	Proportion of bigrams and trigrams which occur in the BNC, i.e. have a frequency > 0.

3. Results

We first compared groups across dependent variables, carrying out one-way ANOVAs unless data were not normally distributed in at least one group (determined by Shapiro-Wilk tests), in which case we carried out Kruskal-Wallis tests. We made six post-hoc pairwise comparisons for each variable and Bonferroni adjustments suggest a significance threshold of p = .008. We tested three variables which indicate formulaicity in combinations: bigram frequency, bigram t-scores and BiBNC ratio. Because of these multiple variables, further correction for multiple comparisons is in order. However, these variables are conceptually related and all based on raw frequencies and in such a case, simple Bonferroni corrections are inappropriate as they inflate the risk of Type II error (McKenzie, 2012; Perneger, 1998). We therefore used modified Bonferroni corrections on the basis of correlations between frequency-related variables (see 3.2 for details). Where results were significant under non-corrected thresholds, or close to thresholds, we noted that observation.

We had directional hypotheses regarding formulaicity values for all groups, as well as for other differences between aphasic groups on one side, and controls and speakers with RHD on the other side (see end of introduction for hypotheses). For pairwise comparisons we report one-tailed p values when hypotheses were directional, and state it in the test.

3.1 Summaries

Table 3 contains a summary group averages and SDs, group effects and significant pairwise comparisons. There were no significant differences between NC and RHD groups, while other groups differed in several variables.

For all results and NC and RHD groups produced more words than both aphasia groups, and speakers with fA produced more than speakers with nfA (p < .001, one-tailed). Participants with RHD produced more words than NC participants, and while the difference was not significant, it was close to the (uncorrected) significance threshold (p < .067, two-tailed). Combination ratio was higher in the NC and RHD groups than in the aphasia groups, and higher in the fA than nfA group (p < .003, one-tailed). TTRs for words, bigrams and trigrams were highest in NCs and lowest in speakers with nfA. Differences were only significant for words and bigrams, where nfA samples had significantly lower TTR than other samples (p < .005, one-tailed).

Content word ratio was highest in the nfA group, and lowest in the fA group. Only the difference between the aphasic groups was significant (p < .001, one-tailed). The difference between the nfA and NC groups was significant under unadjusted thresholds (p = .017, one-tailed), as was the difference between RHD and fA groups (p = .045, two-tailed). The frequency of content words was higher in the aphasic than in NC and RHD groups (p < .001, one-tailed).

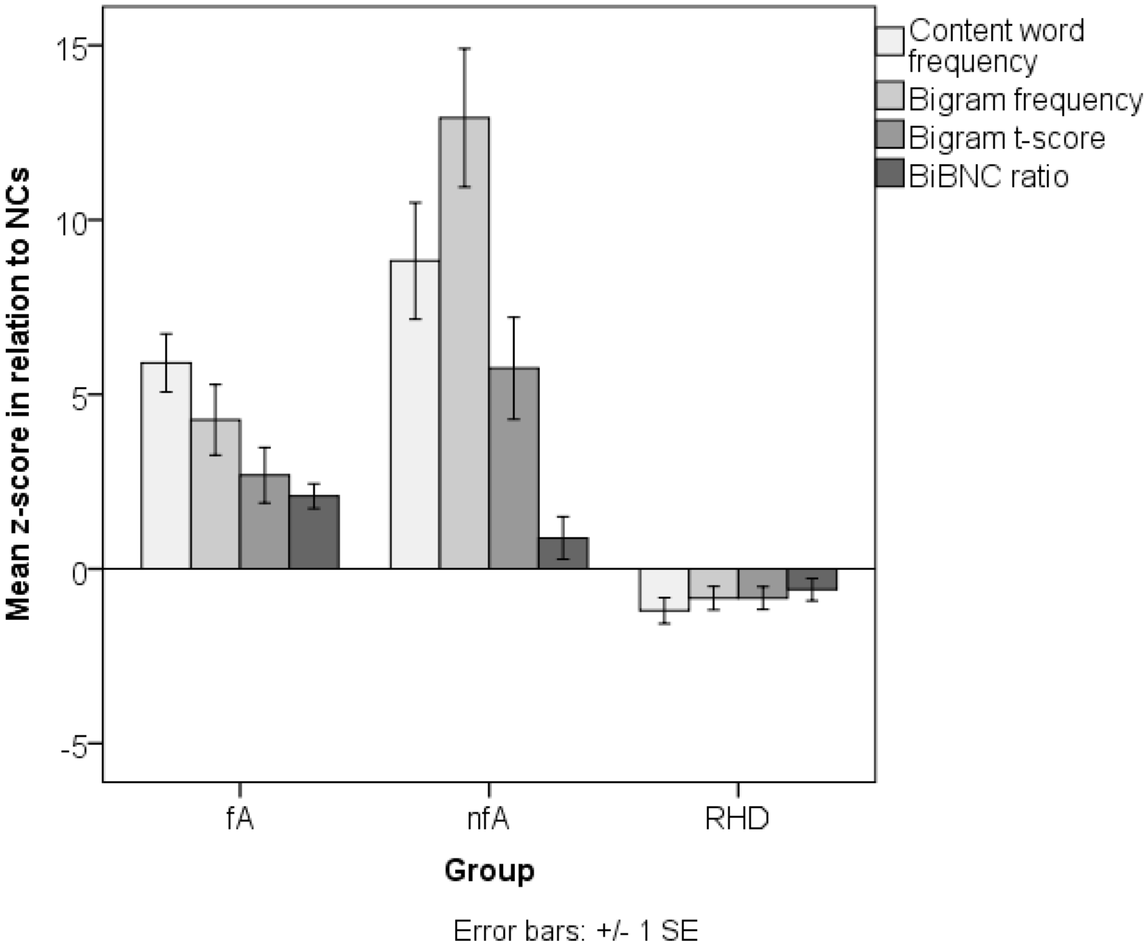
As stated above, combination frequencies, t-scores and BNC ratios are conceptually related and contain the same underlying values. To correct for multiple comparisons we used a solution that takes into account these relationships (McKenzie, 2012; Sankoh, Huque, & Dubey, 1997). We adjusted p values not using the number of comparisons *n*, but rather $n^{1-r(.k)}$, where r(.k) is the average correlation between the outcomes. We ran correlational analyses separately for bigrams and trigrams and across all groups. Correlations were one-tailed since we expected positive correlations throughout. For significance threshold adjustment, only correlation coefficients matter. The mean coefficient for bigrams was r = .641 (frequency and t-scores: r = .923, p < .001; frequency and BiBNC ratio: r = .461, p = .001; t-scores and BiBNC ratio: r = .539, p < .001). For main effects, we therefore adjusted the significance threshold to p = .034 ($.05 / 3^{1-.641}$). For post-hoc pairwise comparisons, we also accounted for the six group comparisons. The modified threshold was p = ($.05 / (6 * 3^{1-.641})$) = .005.

Figure 1 and Table 3 show the results for frequency and frequency-based variables. The average bigram type frequency was significantly higher in participants with nfA than in any other group (p < .001, one-tailed). Speakers with fA produced significantly more frequent bigrams than speakers with RHD (p = .001, one-tailed). The difference between the NC and fA groups was significant only under the uncorrected threshold (p = .006, one-tailed).

Bigram t-scores (collocation strength) in nfA samples was significantly higher than in NC (p < .001, one-tailed) and RHD (p < .001, one-tailed) samples. Differences between the fA group and NC (p = .018, one-tailed) and RHD (.009, one-tailed) groups were significant only under the uncorrected threshold. The average BiBNC ratio was highest in aphasic groups, and lowest in the NC and RHD groups. It was significantly higher in the fA than in NC (p = .001, one-tailed) and RHD (p < .001, one-tailed) groups. Differences between nfA and fA (p = .045, two-tailed) and RHD (p = .008, one-tailed) groups were significant only under the uncorrected threshold.

[INSERT FIGURE 1 HERE]

Figure 1. Content word frequency, bigram frequency, t-scores and BiBNC ratio across all groups of speakers with neurological damage. To allow better comparison between variables which either use different scales or occupy different ranges of the same scale, all values were converted to z-scores using the NC group's mean and SD for the respective variable. Zero therefore represents the control group mean for the variable. Higher values imply more common words and more common/formulaic combinations.



3.2 Post-hoc analyses

3.2.1 Word and combination diversity

After addressing our hypotheses we carried out post-hoc analyses to further interrogate the data. One striking result was that people with nfA had the highest TTRs for words, suggesting increased lexical variety. TTR as a measure of lexical variation is confounded when sample size varies between speakers, causing variation to be overestimated in samples with fewer words (Harris Wright, Silverman, & Newhoff, 2003). In our study, aphasic groups produced such smaller samples, and samples from speakers with nfA were the smallest. This issue can be addressed by using a different measure, namely vocabulary diversity (*D*; McKee, Malvern, & Richards, 2000). *D* is generated by plotting TTR from token samples of increasing size against the size of the random sample, and finding the best fit for the curve. High values indicate greater variety. *D* is not a function of sample size and should therefore be preferred when sample sizes differ between groups. We used the *vocd* function of CLAN (MacWhinney, 2000, version 11/12/2007) to determine *D* for each individual sample (Table 3). Average *D* was significantly higher in NC and RHD groups than in aphasia groups and higher in the fA than in the nfA group ($p < .006$, two-tailed).

Difficulties with TTR extend to bigram and trigram tokens. Because there is no program for calculating *D* for word combinations, we conducted two ANCOVAs using group as independent variable, bigram and trigram TTRs as dependent variables and bigram and trigram count (respectively) as covariates. For bigrams, the effect of group was significant, $F(3,35) = 5.197$, $p = .004$, $\eta_p^2 = .31$. Posthoc comparisons with Bonferroni adjusted significance thresholds (six comparisons; $p = .008$) showed that the nfA group had lower bigram TTR than the other groups ($p < .004$). Differences between the speakers with fA and NCs, as well as between speakers with fA and RHD, were close to the adjusted threshold ($p = .012$) in both comparisons. The covariate (bigram count) also had a significant effect (p =

.001). For trigrams, the effect of group was not significant. F(3,35) = 2.266, p = .098, η_p^2 = .16. The covariate (trigram count) also had no significant effect, p = .606.

[INSERT TABLE 3 HERE]

Table 3. Overview of all critical comparisons (group means and inferential comparisons). Group effects were computed using ANOVAs when distributions within all groups were parametric, and Kruskal-Wallis tests when one or more distributions were non-parametric. The exception are tests for bigram and trigram TTR which were computed using ANCOVAs and bigram and trigram counts (respectively) as covariates. Pairwise differences are only listed if significant after adjustment for multiple comparisons (see 3. for criteria and effects which were close to thresholds).

	NC (SD)	fA (SD)	nfA (SD)	RHD (SD)	Group effect	Significant differences
Word count	3651 (913)	2419 (686)	829 (327)	4232 (687)	F(3,36) = 48.173, $p < .001$, $\eta_p^2 = .8$	NC > fA, nfA RHD > fA, nfA fA > nfA
Combination ratio	.64 (.03)	.49 (.11)	.23 (.13)	.63 (.04)	F(3.36) = 43.722, $p < .001$, $\eta_p^2 = .78$	NC > fA, nfA RHD > fA, nfA fA > nfA
Content word ratio	34.8% (1.7)	33.2% (2.8)	45.2% (11.1)	34.7% (1.1)	$\chi^2(3) = 14.34$, $p = .002$, $\eta_p^2 = .24$	fA < nfA
Content word frequency	311 (27)	473 (72)	553 (145)	278 (140)	F(3,36) = 24.514, $p < .001$, $\eta_p^2 = .67$	NC < fA, nfA RHD < fA, nfA fA < nfA
Bigram frequency	126 (11)	173 (35)	268 (69)	117 (12)	F(3,36) = 30.69, $p < .001$, $\eta_p^2 = .72$	nfA > NC, RHD, fA fA > RHD
Bigram t-score	13.06 (1.17)	16.2 (2.94)	19.78 (5.42)	12.09 (1.2)	F(3.36) = 11.836, $p < .001$, $\eta_p^2 = .5$	nfA > NC, RHD
BiBNC ratio	.9 (.016)	.94 (.022)	.92 (.038)	.89 (.02)	F(3,36) = 7.795, $p < .001$, $\eta_p^2 = .41$	fA > NC, RHD

Word TTR	.19 (.03)	.18 (.04)	.25 (.04)	.18 (.02)	$\chi2(3) = 13.47$, p = .004, $\eta_p^2 = .24$	nfA < NC, RHD, fA
Bigram TTR	.65 (.05)	.64 (.06)	.61 (.09)	.64 (.04)	F(3,35) = 5.197, p = .004, $\eta_p^2 = .31$	nfA < NC, RHD, fA
Trigram TTR	.9 (.03)	.88 (.03)	.77 (.16)	.89 (.03)	F(3,35) = 2.266, p = .098, $\eta_p^2 = .16$	-
Vocabulary diversity (*D*)	104 (16.27)	78 (11.75)	32 (15.51)	95 (6.65)	F(3,36) = 59.078, p < .001, $\eta_p^2 = .83$	NC > fA, nfA RHD > fA, nfA fA > nfA

3.2.2 Reliability of the BNC as a reference corpus

To test the degree to which the spoken BNC can be relied upon as a reference corpus for the current study, we correlated BNC frequencies for words and bigrams types with the frequency in which the respective unit occurred in our NC test sample. We interpreted high correlations as evidence that frequency patterns in the reference corpus represents natural language production in our populations. 2706 word types in the NC sample occur in the spoken BNC. The correlation between frequency in the BNC and frequency in the NC sample was large, r = .924, p < .001. Of the bigram types in the NC sample, 4667 occur in the spoken BNC. The correlation between frequency in the BNC and frequency in the NC sample was large, r = .809, p < .001.

3.2.3 Most frequent words and combinations within samples

To gain further insight into the production of words and word combinations, we extracted and explored the ten most frequently used word and trigram types within each group's sample, along with the proportion of tokens represented by each type (Tables 4 and 5). We also made some more qualitative observations. Unsurprisingly, the list consists exclusively of function words. The list of most frequent words is very similar across NC, fA and RHD groups, sharing eight of the ten words (*I*; *and*; *the*; *to*; *it*; *a*; *you*; *that*). Only three of these words (*I*; *and*; *it*) are among the most frequent words in the nfA sample. While the most common words in the nfA sample are predominantly items that can serve as discourse markers or agreement tokens, such as *yes*, *yeah*, *aye*, *no*, *oh*, *and* and *but*, words in the other lists are more associated with phrasal syntactic functions. Differences in lexical diversity also manifest in this analysis. The proportion of overall tokens represented by the ten most frequent types is almost similar between NC and RHD samples, relatively higher in the fA sample and markedly higher in the nfA sample.

[INSERT TABLE 4 HERE]

Table 4. The ten most frequent words within each group, and the proportions of word tokens each type represents. The total proportion for all ten types is displayed under the group label.

NC (total: 26.5%)	fA (total: 27.3%)	nfA (total: 36.1%)	RHD (total: 26.3%)
I (5.6%)	I (7.1%)	no (5.7%)	I (7.1%)
and (4.1%)	and (3.4%)	and (5.4%)	and (3.5%)
the (3.1%)	it (2.8%)	yes (4.6%)	to (2.5%)
to (2.5%)	to (2.3%)	I (4.3%)	it (2.3%)
it (2.4%)	that (2.2%)	aye (3.1%)	the (2.1%)
a (2.1%)	the (2.1%)	yeah (3%)	a (2%)
you (1.9%)	n’t (2%)	it (2.8%)	you (1.9%)
that (1.8%)	a (1.9%)	oh (2.6%)	that (1.7%)
‘s (1.6%)	you (1.9%)	‘s (2.4%)	‘s (1.6%)
n’t (1.5%)	was (1.8%)	but (2.3%)	in (1.6%)

The difference between the nfA and other samples is even greater at the trigram level. The ten most frequent trigram types accounted for 1.6%, 2.1% and 3.5% respectively within NC, RHD and fA samples. This proportion was 10.6% in the nfA sample. There are fewer common items across lists than at word level, but it is striking that *I don’t* and *don’t know* occurs on all lists, and are respectively the most and second most common trigram in NC and aphasic samples. Despite nfA being associated with paucity of function words, function words clusters dominate the list.

[INSERT TABLE 5 HERE]

Table 5. The ten most frequent trigram types within each group and the proportions of trigram tokens each type represents. The total proportion for all ten types is displayed under the group label.

NC (total: 1.6%)	fA (total: 3.5%)	nfA (total: 10.6%)	RHD (total: 2.1%)
I don’t (.4%)	I don’t (.8%)	I don’t (2.9%)	I don’t (.4%)
don’t know (.2%)	don’t know (.5%)	don’t know (2.5%)	I’ve got (.3%)
I didn’t (.2%)	I didn’t (.5%)	I can’t (1.6%)	I can’t (.2%)
a lot of (.2%)	I can’t (.4%)	it’s a (1.1%)	as I say (.2%)
I can’t (.2%)	that’s right (.3%)	but it’s (.6%)	I couldn’t (.2%)
I’ve been (.1%)	I couldn’t (.3%)	can’t say (.5%)	don’t know (.2%)
didn’t know (.1%)	I used to (.2%)	I used to (.4%)	I’ve been (.2%)
and I was (.1%)	and then I (.2%)	two or three (.4%)	I used to (.2%)
I used to (.1%)	I’m not (.2%)	but I can (.3%)	a lot of (.2%)
it’s a (.1%)	I went to (.2%)	n’t say it (.3%)	you’ve got (.1%)

4. *Discussion*

This study set out to analyse spontaneous language from speakers with aphasia, RHD and neurotypical controls on the basis of frequency of usage and other quantitative measures. We used the FLAT, a computer script that counts word and combination tokens and types, looks up their frequency in a normative corpus and uses the data to determine expressive range and formulaicity of speakers. In a previous study (Zimmerer et al., 2016), FLAT values distinguished between neurotypical speakers and people with Alzheimer's disease, a condition associated with an increase in formulaic language. Values also correlated with estimated time post-symptom-onset. In a previous investigation, Van Lancker Sidtis and Postman (2006) used raters to categorize utterances, or part of utterances, as formulaic or non-formulaic. They found that aphasic speakers were more formulaic than healthy controls, while speakers with RHD were less formulaic.

Our FLAT data overall support the previous findings for people with aphasia. We start the summary with more general measures. During the half hour interview, speakers in each aphasic group produced fewer words than NCs, and nfA speakers produced fewer words than any of the groups. Patterns were similar for fluency: Speakers with aphasia produced proportionally fewer combinations than the NC group, and speakers in the nfA group produced fewer than any other group. Our use of TTR to measure diversity of expression was flawed, since TTRs are a function of sample length, which differed significantly between groups. In post-hoc analyses we used *D* to investigate diversity at word level, and entered bigram and trigram counts as covariates into comparison of bigram and trigram TTR. People with aphasia showed less lexical diversity than controls, with the nfA group showing the least. These results are similar to previous findings by Harris Wright, Silverman and Newhoff (2003). For combinations, we only found a significant difference between participants with nfA and NCs, and only for bigrams. However, we cannot determine the degree to which a comparison of word combination TTR understates the difference between aphasic and other groups even after entering the covariate. Previously, Perkins (1994) found repetitiveness in multiword expressions of two individuals with fA. In a post-hoc analysis we found that speakers with nfA in particular showed a narrower distribution of words and trigrams, with the ten most frequent types of each accounting for a larger proportion of tokens than other groups, further illustrating loss of expressiveness.

Both aphasic groups produced significantly more frequent content words than NCs, and both differed from controls at the bigram level. With regards to the statistical significance of bigram effects, we observed that fA and nfA groups had somewhat mirrored profiles in relation to NCs. For a given bigram variable in which one group showed a significant difference to NCs, the other group showed a trend that was very strong, but not significant at Bonferroni corrected levels. Going by effect sizes as indicated by z-scores (Figure 1), speakers with fA differ from NCs most at the single word level, while speakers with nfA differ most at the bigram level. Speakers with nfA are more likely to produce a “rare” bigram, i.e. a bigram with a frequency of zero in the spoken BNC. At the same time, when they produce bigrams that are not rare, they are more frequent and have greater collocation strength. Given the limited sample size of ten speakers per group, replication is required for confirmation and further exploration. However, results seem to underline the different neurological and linguistic nature of fluent and non-fluent aphasic speakers, and in particular, views that fA is a more lexical, and nfA a more grammatical impairment.

None of our measures distinguished speakers with RHD from NCs at significant levels. However, frequency and collocation effects did go in the hypothesized direction. In light of positive findings of Van Lancker Sidtis and Postman (2006), we discuss two explanations for our results. First, given smaller effect sizes in comparison with aphasic groups, more statistical power, i.e. participants, may be needed. Note that Van Lancker Sidtis and Postman also found relatively small effect sizes for the RHD group. Second, the inclusion criterion for our group was damage to any part of the right hemisphere, which likely resulted in the group being more heterogeneous than the aphasic group, which by definition had damage to areas critical for language. Future studies may focus on participants with damage to particular areas of the right hemisphere, importantly frontotemporal areas such as Broca's and Wernicke's area homologues.

Results suggest that to produce grammatical utterances, speakers with aphasia rely on more frequent and strongly collocated combinations, which we regard as more likely to be holistically processed, i.e. formulaic. This is likely the result of impaired networks responsible for more creative, combinatorial processes. Grammatical errors, which are an important feature of aphasic language but were not our focus, are another source of evidence for impaired syntax. A complete picture of aphasic language needs to take into account properties of grammatical and ungrammatical production, and we propose that the approach based on usage-frequency makes a valuable contribution to profiling and understand language impairment.

We consider the strengths of our study to be the novel approach to aphasic language evaluation, using an automated, blind analysis, and the size of individual transcripts. Limitations are the possibly heterogeneous RHD group and the number of participants within groups. Further, because this is a sample from an older study, we cannot provide a full psycholinguistic profile using contemporary language test batteries. Our vocabulary diversity measures (*D*) are very similar to Harris Wright et al.'s (2003) results, which suggests that our aphasic samples may be typical for studies of this type, but the question remains as to how severity of impairment, as measured by more established tests, affected the results. A preliminary report (Zimmerer, Coleman, Hinzen, & Varley, 2017) suggests that FLAT values, in particular combination frequency, are related to performance in sentence comprehension tasks since aphasic participants with weaker comprehension performance used fewer rare combinations.

FLAT analysis does not allow a qualitative description of formulas, for instance a categorization by type of formula. Van Lancker Sidtis and Postman's (2006) rater-based analysis found a change in the distribution of formulas across NC, aphasic and RHD groups. People with aphasia produced proportionally more discourse elements than NCs, and fewer formulas containing proper nouns. In speakers with RHD there was a reduction of discourse elements and pause fillers. This difference may contribute to the impression of RHD language as "disinhibited" (Bates et al., 2001). Our post-hoc analysis suggests that the increased use of discourse elements in aphasic individuals may be driven by speakers with nfA, as their most frequent word types were discourse particles. The most frequent words in the other groups, including fA speakers, were predominantly associated with syntactic functions. Our exploration of the most used trigrams in each sample also suggests that some change in fA and nfA output can be attributed to increased reliance on a smaller group of function word clusters. Results demonstrate that the disruption of function words is not absolute – rather, use of function words appears more dependent on available formulas.

FLAT looks at the frequency of word and word combinations in their particular form, with the exact affixation (and in some cases contraction) as transcribed. It is likely that frequency effects also apply at other levels of representation, i.e. at the level of lemma, argument structure type, or voice. As one reviewer of our manuscript suggests, it would be interesting to investigate frequency of a lemmatised transcript, or of a transcript in which grammatical errors were correct by the transcriber as a mean to move away from the representation at surface form.

Our outcomes and interpretation are consistent with theoretical frameworks such as construction grammar that see formulaic language, as represented by holistic multi-word chunking (Christiansen & Chater, 2017; Goldberg, 2003), as a core aspect of language. While these frameworks have had growing influence in other fields such as language acquisition (e.g., Tomasello, 2003), they have rarely been applied to aphasia and other forms of neuropathology (Gahl & Menn, 2016). One important future direction is to determine the explanatory power of usage-based frameworks in direct comparison to the various "words and rules" approaches that have emerged to explain grammar in aphasia (Bastiaanse & van Zonneveld, 2006; Beretta & Campbell, 2001; Grodzinsky, 2000). While we present our data in light of usage-based theories, we have yet to understand the contribution of type of grammatical construction, since some constructions assumed to be computationally simple (e.g., actives) are more frequent than difficult constructions (e.g., passives).

It will take further research to determine reliable normative values for neurotypical populations, taking into account important variables such as age, education, profession and dialect. Currently, we also do not know which language elicitation task is most valuable for this purpose (e.g., conversations, picture descriptions, elicitation of personal narratives such as an account of a recent holiday), and how long samples should be for detection of subtle language change as can occur in early-stage dementia or might follow a therapeutic intervention. The choice of reference corpus also needs to be considered, taking into account corpus size and text style, since word and combination distributions vary substantially across corpora (Conrad & Biber, 2004; Gray & Biber, 2013; Stubbs & Barth, 2003), which means that the choice of "best corpus" can depend on the nature of language elicitation. Corpora also differ in the extent to which frequency interacts with other language properties. For example, in subtitle corpora frequency is more strongly correlated with emotional values, likely as the result of script writers' tendency to write more emotional dialogue (Baayen et al., 2016). The BNC may be a good compromise for researchers working with different types of spoken output, and our own test in section 3.3.2 suggests that it was reliable for our study.

Our results show that frequency properties of language output can change substantially as the results of aphasia. They demonstrate the potential of computerized and corpus based analyses in contributing to clinical diagnosis or tracking of change over time, for example as a response to intervention. In comparison to other approaches these analyses can be applied by researchers and clinicians with minimal linguistic training.

Authorship statement and acknowledgements

V.Z. and R.V. conceived the study. R.V. acquired the data. L.N. and R.T. transcribed interviews and formatted them for analysis. V.Z. designed the computerised analysis, M.C. wrote the programme. V.Z., L.N. and R.T. analysed the data. V.Z. wrote the first draft of the manuscript. R.V., V.Z. and L.N. contributed to further drafts. We thank Lucy Ferguson for her assistance with transcription and formatting and Claudia Bruns for suggestions and comments. We also thank Kathy Conklin and one anonymous colleague for reviewing the manuscript.

The authors have no financial conflicts of interest.

References

Anthony, L. (2018). AntConc. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. http://doi.org/10.1016/j.jml.2009.09.005

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220. http://doi.org/10.1080/02687038.2016.1147767

Bastiaanse, R., & van Zonneveld, R. (2006). Comprehension of passives in Broca’s aphasia. *Brain and Language*, *96*(2), 135–142. http://doi.org/10.1016/j.bandl.2005.06.012

Bates, E. B., Reilly, J., Wulfeck, B., Dronkers, N., Opie, M., Fenson, J., … Herbst, K. (2001).

Differential effects of unilateral lesions of language production in children and adults. *Brain and Language*, *79*, 223–265. http://doi.org/10.1006/brln.2001.2482

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585–594. http://doi.org/10.1001/archneur.1994.00540180063015

Beretta, A., & Campbell, C. (2001). Psychological verbs and the double-dependency hypothesis. *Brain and Cognition*, *46*(1–2), 42–6. http://doi.org/10.1016/S0278-2626(01)80030-0

Blanken, G. (1991). The functional basis of speech automatisms (recurring utterances). *Aphasiology*, *5*(2), 103–127. http://doi.org/10.1080/02687039108249477

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

Christiansen, M. H., & Chater, N. (2017). *Creating Language. Intergrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

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.

Code, C. (1983). On "neurolinguistic analysis of recurrent utterances in aphasia": Reply to de Bleser and Poeck. *Cortex, 19*, 261–264.

Code, C. (1989). Speech automatisms and recurring utterances. In C. Code (Ed.), *The*

Characteristics of Aphasia (pp. 155–163). Hove and London: Laurence Erlbaum.

Code, C. (2005). First in, last out?: The evolution of aphasic lexical speech automatisms to agrammatism and the evolution of human communication.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*, 72–89. http://doi.org/10.1093/applin/amm022

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. http://doi.org/10.1017/S0267190512000074

Conrad, S. M., & Biber, B. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*, 56–71. http://doi.org/10.1515/9783484604674.56

Cuetos, F., Arce, N., Martínez, C., & Ellis, A. W. (2015). Word recognition in Alzheimer's disease: Effects of semantic degeneration. *Journal of Neuropsychology*. http://doi.org/10.1111/jnp.12077

Cuetos, F., Rodríguez-Ferreiro, J., Sage, K., & Ellis, A. W. (2012). A fresh look at the predictors of naming accuracy and errors in Alzheimer's disease. *Journal of Neuropsychology*, *6*(2), 242–56. http://doi.org/10.1111/j.1748-6653.2011.02025.x

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349. http://doi.org/10.1080/01690969008407066

f4transkript. (2012). Marburg, Germany: audiotranskription.de.

Gahl, S., & Menn, L. (2016). Usage-based approaches to aphasia. *Aphasiology*, 1–17.

http://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. http://doi.org/10.1016/S1364-6613(03)00080-9

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–136. http://doi.org/10.1075/ijcl.18.1.08gra

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. http://doi.org/10.1017/S0140525X00002399

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, *89*(1), 151–174. http://doi.org/10.1111/j.2044-8295.1998.tb02677.x

Harris Wright, H., Silverman, S., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology*, *17*(5), 443–452. http://doi.org/10.1080/02687030344000166

Jacobs, C. L., Dell, G. S., & Bannard, C. (2017). Phrase frequency effects in free recall: Evidence for redintegration. *Journal of Memory and Language*, *97*, 1–16. http://doi.org/10.1016/J.JML.2017.07.003

Jaecks, P., Hielscher-Fastabend, M., & Stenneken, P. (2012). Diagnosing residual aphasia using spontaneous speech analysis. *Aphasiology*, *26*(7), 953–970. http://doi.org/10.1080/02687038.2012.663075

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PloS One*, *7*(3), e33202. http://doi.org/10.1371/journal.pone.0033202

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843. http://doi.org/10.1037/0278-7393.20.4.824

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Manwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2007). TalkBank Project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora: Synchronic Databases, Vol. 1*. Houndmills: Palgrave-Macmillan.

McKee, G., Malvern, D. D., & Richards, B. J. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–338. http://doi.org/10.1093/llc/15.3.323

McKenzie, D. (2012). Tools of the trade: A quick adjustment for multiple hypothesis testing. Retrieved February 8, 2017, from http://blogs.worldbank.org/impactevaluations/tools-of-the-trade-a-quick-adjustment-for-multiple-hypothesis-testing

Perkins, M. (1994). Repetitiveness in language disorder: A new analytical procedure. *Clinical Linguistics & Phonetics*, *8*(4), 321–336.

Perneger, T. V. (1998). What’s wrong with Bonferroni adjustments. *BMJ*, *316*(7139). http://doi.org/10.1136/bmj.316.7139.1236

Sankoh, A. J., Huque, M. F., & Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, *16*(22), 2529–42. http://doi.org/10.1002/(SICI)1097-0258(19971130)16:223.0.CO;2-J

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. http://doi.org/10.1016/j.bandl.2017.10.004

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators. *Functions of Language*, *10*(1), 61–104. http://doi.org/10.1075/fol.10.1.04stu

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

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

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

Van Lancker Sidtis, D., Choi, J., Alken, A., & Sidtis, J. J. (2015). Formulaic language in Parkinson's disease and Alzheimer's disease: Complementary effects of subcortical and cortical dysfunction. *Journal of Speech, Language, and Hearing Research : JSLHR*, *58*(5), 1493–507. http://doi.org/10.1044/2015_JSLHR-L-14-0341

Van Lancker Sidtis, D., & Postman, W. A. (2006). Formulaic expressions in spontaneous speech of left- and right-hemisphere-damaged subjects. *Aphasiology*, *20*(5), 411–426.

Varley, R. (1993). Deictic terms, lexical retrieval and utterance length in aphasia: An investigation of inter-relations. *International Journal of Language & Communication Disorders*, *28*(1), 23–41. http://doi.org/10.3109/13682829309033141

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

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. http://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. http://doi.org/10.1017/S026719051200013X

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: an integrated model. *Language & Communication*, *20*, 1–28. http://doi.org/10.1016/S0271-5309(99)00015-4

Zimmerer, V. C., Coleman, M., Hinzen, W., & Varley, R. A. (2017). Reliance on common word combinations correlates with degree of syntactic impairment in aphasia. *Stem-Spraak- En Taalpathologie*, *22*(Supplement 2), 163.

Zimmerer, V. 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. http://doi.org/10.3233/JAD-160099