1

**Zimmerer, V. C., Wibrow, M., Varley, R. A. (accepted). Formulaic language in people with probable Alzheimer's Disease: a frequency-based approach.** *Journal of Alzheimer's Disease*.

# Formulaic language in people with probable Alzheimer's Disease: a frequency-based approach

Vitor C. Zimmerer[a], Mark Wibrow[b], Rosemary A. Varley[a]

[a]Department of Language and Cognition, University College London

[b]Department of Speech, Hearing and Phonetic Sciences, University College London


**Correspondence address:**

Vitor Zimmerer

Department of Language and Cognition

University College London

Chandler House

2 Wakefield Street

London WC1N 1PF

Telephone: +44 (0)20 7679 4270

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

Abstract

Background: Language change can be a valuable biological marker of overall cognitive change in Alzheimer's Disease (AD) and other forms of dementia. Previous reports have described increased use of language formulas in AD, i.e. combinations likely processed in a holistic manner. Words that commonly occur together are more likely to become a formula. Objective: To determine if frequency of co-occurrence as one indicator for formulaic language can distinguish people with probable AD from controls and if variables are sensitive to time post-symptom onset. Methods: We developed the Frequency in Language Analysis Tool (FLAT) which indicates degrees of formulaicity in an individual

language sample. The FLAT accomplishes this by comparing individual language samples to co-occurrence data from the British National Corpus (BNC). Our analysis also contained more conventional language variables in order to assess novel contributions of the FLAT. We analysed data from the Pitt Corpus which is part of DementiaBank. Results: Both conventional and co-occurrence variables were able to distinguish AD and control groups. According to co-occurrence data people with probable AD produced more formulaic language than controls. Only co-occurrence variables correlated with disease progression. Discussion: Frequency of word co-occurrences is one indicator for formulaicity and a valuable contribution to characterizing language change in AD.

Keywords: Alzheimer's Disease; Language; Verbal behaviour; Disease progression

1. Introduction

There are two ways to produce sentences. One makes use of a combinatorial system that generates them out of words and affixes which are each retrieved as a separate unit (e.g., {I, do, not, know}). For this reason, the process has been labelled "atomic" [1]. Atomic descriptions of language processing have dominated the language sciences, especially since the ascent of generative "words and rules" models in the mid-20th century [2–4]. "Holistic" approaches on the other hand suggest that some word combinations, be they sentence fragments or even entire utterances, are accessed as a whole. These representations are commonly referred to as formulas [5–7]. Some formulas can contain gaps to be filled (e.g., *What's* X?), while others may be completely specified (e.g., *I don't know*). As a result, sentences may be produced with little or no combinatorial effort. Proponents of holistic models have suggested two criteria for formulaicity: first, idiomatic utterances, such as "X *kicked the bucket*" are said to be formulaic to a certain degree because the correct interpretation (in this example "X died") cannot be derived by combining the meaning of the individual words "kick" and "bucket". Second, formulas result from frequency of co-occurrence in everyday language use. Words that occur together regularly may become holistic [8–11] and high-frequency utterances such as *I don't know* [12] are prime examples. In this report we use frequency data to investigate how changes in the degree of

formulaicity in speech output may be a marker of cognitive change resulting from Alzheimer's Disease (AD).

AD leads to changes in behavior across a range of language domains, from word-retrieval to the richness and cohesion of narrative [13]. Linguistic behavior may provide sensitive markers of disease progression and as a result, research into the language of dementia can inform diagnosis, tracking of change, and measuring the effects of intervention. Language data are abundant, easily acquired, and are of high functional value as they reflect a person's ability to interact with family and peers, consume media and engage in work. Approaches to language in dementia have often looked at the pragmatics of discourse and narrative [14–18] and at lexical processing [19–21], with word frequency predicting lexical production and recognition [22,23].

Grammatical change in AD is subtle, at least in early stages, and contrary to aphasias, language in AD contains only a small (but significant) increase in syntactic errors [24]. By contrast, word-finding difficulty and lexical errors are more evident, leading to the claim that early stage AD affects semantic processing, but not grammatical capacity [25]. The claim of lexical-semantic deterioration in the face of preserved grammatical processing is largely consistent with atomic views of language. In particular, rule-based mechanisms of grammar continue to generate well-structured output, although planned structures are sometimes aborted due to failures of word retrieval from semantic memory. There is, however, a marked reduction of grammatical complexity in AD, and typical features include reduced sentence length, fewer conjunctions, simpler verb phrase structure, and less frequent use of passives, relative clauses and other complex constructions [26–28]. Ahmed, Haigh, Jager and Gerrard [29] examined language samples from people with a diagnosis of AD which was subsequently confirmed at autopsy. They found that the distribution of word classes in language production as well as "syntactic complexity", (a compound variable including mean length of utterance, proportion of words in sentences, number of embedded clauses, syntactic errors, nouns preceded by determiners and

verbs with inflections) distinguished individuals with AD from controls, as well as from speakers with mild cognitive impairment.

Bates et al. [26], in their work on the underlying source of language deviation in AD, discuss the notion of automaticity. They argue that simpler sentence structures are overlearned and can be retained despite linguistic-cognitive impairment, while more complex structures require more explicit processing and inhibition of automatic production. However, while Bates et al. [26] focused on structure type, we extend the notion of overlearned structures to word combinations in general, regardless of the sentence type in which they appear. *I don't know* and *I can't reach* are of the same structural type, but the former is easier to produce because of properties of the specific sequence of words. The former effectively communicates uncertainty and is very useful in everyday communication, in particular for speakers with communication difficulties. *I don't know* is also the more frequent sentence, and its individual words are more strongly associated with each other. Both frequency of co-occurrence and function can predict why one sentence is more formulaic, and therefore cognitively entrenched, than another. In many cases, we can expect these factors to be related. Many words frequently co-occur because, as a combination, they serve an important function, be it pragmatic or grammatical.

While we are considering formulaicity as a disease marker, note that there is nothing pathological about using formulas per se. They are considered part of typical language use, and it has been estimated that one third to one half of healthy discourse is formulaic [7]. Formulas are thought to be crucial for language acquisition, and influential accounts propose that children first learn a range of formulas before being able to  manipulate individual words and affixes [30,31]. Advantages in the processing of formulas extend to adulthood, where high-frequency word combinations are recognized and produced more rapidly than low-frequency combinations [8,32–34]. At the same time, individual words can be harder to recognize if they appear within high-frequency combinations [9], supporting the view that they prioritize processing at holistic and not atomic levels.

Formulas are useful in that they reduce cognitive demand: their meaning is accessed holistically and they require less structural computation. Due to the reduced processing demands of understanding and producing formulas, an increase in their use may be a subtle marker of neurocognitive impairment. Residual formulas are often apparent in severe aphasia [35–37]. Zimmerer, Cowell and Varley [38] describe residual formula use in severe aphasia, where speakers continue to employ phrases such as *I don't know*; *and then,* in the absence of ability to modify these combinations to produce, for example, *I know*, *You don't know* or *I don't think*. An increase in "common phrases" has also been observed in AD [39]. Wray [40] argued that overly formulaic language in AD is a considerable barrier between patient and conversation partner, locking both into repeating patterns of behavior. An increase in formulaicity comes with a decrease of combinatorial, creative power and the ability to communicate a wider range of opinions, ideas and needs.

One methodological approach has been applied consistently in a series of studies on formulaic language in pathology. Developed by Lancker-Sidtis and Rallon [41], the methods involve determining the proportion of words which independent raters consider part of a formulaic expression. Formulas are defined as familiar word selections and combinations with stereotyped intonation. Using this method, van Lancker-Sidtis and Postman [42] investigated formulaicity in speakers with aphasia following left-hemisphere damage, speakers with right hemisphere damage, and neurotypical controls. People with aphasia were more formulaic than controls, while people with right hemisphere damage appeared less formulaic. In a later study, people with subcortical damage were also found to produce fewer formulas [43]. The same methods revealed increased formulaicity in people diagnosed with AD, not only compared to healthy controls [44], but also to people with Parkinson's Disease. The latter group also displayed reduced formulaicity as the result of the disease [45]. The methods developed by van Lancker Sidtis and colleagues allowed new insights into the language of dementia, and suggest that evaluation of formulaic language allows identification and differentiation between pathologies.

However, systematic research on formulacity is still in its infancy. We explore a new approach which differs from the above in a number of aspects. Methods were designed for language research, but also to support future clinical practice in differential diagnosis or tracking change, for instance in the context of intervention research. With regards to cost effectiveness, the development of automated tools to support clinical assessment yields benefits in terms of speed of data processing, advantages of blinded evaluation and no, or little, need for specialist linguistic training. Our approach addresses these needs.

The most important difference is our focus on frequency of co-occurrence. While frequency, as we laid out, is not all there is to formulaic language, frequency values can be extracted automatically to achieve an estimate of the degree of formulaicity in individual speakers and groups. Further, we abandon a binary classification, where a given utterance is either rated as a formula or not. Language exposure and use differs from individual to individual, and in addition we cannot suggest a specific threshold beyond which a word combination should be considered a formula. Within our approach it is therefore difficult to label with certainty a particular combination in a particular speaker as a formula. Instead, we adopt the view that frequency of co-occurrence in a word combination indicates a degree of formulaicity which indicates how strongly words are associated with each other. As a final innovation, our study evaluates the power of frequency measures by comparing it to a range of other language variables.

Our methods are a variation of the computerized "Traceback" method [31,46,47] which divides large speech samples from individuals (about 30 hours per speaker) into a reference corpus and a smaller test corpus, and determines the extent to which utterances from the test corpus can be derived either verbatim or with small replacements (such as a new noun phrase) from the reference corpus. The results indicate how much a speaker "recycles" utterances, and therefore can provide measures of formulaicity. To our knowledge, Traceback has to date not been applied in a pathological context and

the recording and transcription times needed to acquire sufficient sample sizes make it problematic for clinical use.

For use with smaller samples more typically available from patients, we present the Frequency in Language Analysis Tool (*FLAT*) [48]. Instead of using the speaker's own production as a reference corpus in order to establish the recycling-rate, the FLAT calculates the degree to which the speaker produces word combinations that are common in typical language use. For reference, it uses the BNC, a 100-million-word-corpus containing written and transcribed spoken text from different contexts. The spoken BNC corpus forms a subset of ca. 10 million words. While the BNC cannot account for individual language experience and the resulting differences in formulaic language representation, it is a sizeable resource can be used as a reference corpus for smaller test corpora. Word combinations with high co-occurrence values can be seen as common and therefore more likely to be processed as a formula. The FLAT produces raw frequencies (e.g., the most frequent full finite verb phrase is *I don't know*, occurring 117 times in every million words), but also t-scores which indicate how strongly words are associated with each other considering their individual frequency. Differences in average t-scores are more likely to be the result of how speakers combine words, and not of which words they retrieve (see 2.2.2).

Input to the FLAT is any orthographical text, such as transcribed speech. From each text, the FLAT extracts all words, combinations of two words (bigrams), and combinations of three words (trigrams), and retrieves  their occurrences in the spoken portion of the BNC. It extracts overall frequencies and uses them to compute co-occurrence values. Means for each language sample indicate the "degree of formulaicity". The generation of these variables is automated. This step is therefore blind to any potential researcher bias. The FLAT is also fast, processing hundreds of data points per minute.

For the current study, we used the FLAT to investigate language output in people with probable AD and neurotypical controls. We analyzed participant data from the Pitt Corpus, which is part of the DementiaBank section of the TalkBank project [49]. Participants visited the University of Pittsburgh's

School of Medicine several times over a period of years as part of a dementia study. First visits took place between 1983 and 1988. We analyzed data from the Boston Cookie Theft picture description task [50], which is commonly used to elicit connected speech samples. In this task, participants describe a drawing showing a complex kitchen scene. In contrast to spontaneous production or interviews, the Cookie Theft task allows control for content as well as talk situation. We analyzed the earliest Cookie Theft description available for each speaker to avoid effects of repeated testing. Samples were heterogeneous enough to investigate the relationship between co-occurrence variables and age, education, estimated onset of AD and performance in the Mini Mental State Examination [51]. We focused our analysis on grammatical word combinations, i.e. syntactically "successful" word combinations in English. Based on previous reports, we hypothesized these combinations would be more formulaic in the AD group. We also expected positive correlations between co-occurrence values and estimated time post-onset, and negative correlations between the former and MMSE scores.

In order to determine the value of co-occurrence variables in examining language in neuropathology, we also investigated their relationship with other language measures. We chose variables that we predicted would correlate with co-occurrence variables. We calculated the number of words produced during descriptions, rates of grammatical and semantic errors, the rates of closed-class words and weakly inflected words in each sample, and a measure of connected speech, i.e., how many words are parts of word combinations (see Box 1 for a summary of the variables). Within "words and rules" approaches these measures can be seen as indicating the efficiency of retrieving appropriate lexical items and applying grammatical knowledge or rules. We therefore set up these variables to compete with our frequency-based values to investigate the latter's contribution. Note however that variables like number of words and rate of closed-class words cannot be disentangled from holistic frameworks. Access to formulas may influence verbosity, and it has been noted that many formulaic utterances consist of clusters of closed-class words [52]. We elaborate on these relationships in the discussion.

2.  Materials and Methods

2.1 Participants

The Pitt Corpus lists 510 speakers, who either had a diagnosis of AD or were neurotypical controls. For our study we excluded: 1) Participants for whom Cookie Theft data from the first visit were not available. 2) People with an AD diagnosis who had other pathological diagnoses at the first or subsequent visits (e.g., anxiety, depression, cerebrovascular disease, or other types of neurodegeneration). 3) Controls with a pathological diagnosis at the first or subsequent visits. We excluded samples on the basis of later diagnoses due to the possibility of language change in prodromal states. After applying the selection criteria, we further excluded one sample (087-0) which was allocated the wrong media file and transcript in the database (the issue has been confirmed by a member of the TalkBank project; Davida Fromm, personal communication).

The final sample included data from 48 people with a diagnosis of AD (36 female; mean age = 71.5, SD = 9, range 56-88; mean number of years of formal education = 11.6, SD = 2.6, range 6-17; and 38 healthy controls (23 female; mean age = 62.3, SD = 8.6, range 47-80; mean number of years of formal education = 14.1, SD = 2.2; range 11-20). Mean MMSE score for the AD group was 18.7, SD = 4.3, range 10-27, while the mean score for the controls was 29.2, SD = .8, range 27-30. For the AD group the mean number of estimated years post-onset was 4, SD = 2.45, range 0-12, with a mean estimated age of onset of 67.5, SD = 8.5, range 53-85. Both time post-onset and age of onset were estimated on the basis of the individual's clinical history. Groups differed significantly across several variables. People with probable AD were older, $t(84) = 4.753$, $p < .001$, $d = 1.07$, $r = .47$, had fewer years of formal education, $U = 1377.5$, $p < .001$, $r = .45$, and had lower MMSE scores, $U = 1823$, $p < .001$, $r = .86$. Differences in education may reflect a general tendency of neurotypical people with more years of education to volunteer for research studies. In order to maximize sample size, we decided not to match groups on age and education. To account for these differences, we entered both variables into our models as covariates and examined whether other differences survive residualization. We used a

ranked analysis of covariance [53] to account for non-parametric distributions. We residualized ranked dependent variables over the ranked covariates (in this case age and years of education) and used independent samples t-tests to compare the unstandardized residuals. Differences in MMSE scores were still significant and strong after residualization, $t(84) = 15.836$, $p < .001$; $d = 4.15$, $r = .9$.

## 2.2 Procedure

### 2.2.1    Preparation of transcripts

Cookie Theft transcripts were formatted for analysis with the FLAT. In the first step, we excluded the CHILDES annotation layer and selected only the orthographic transcription and speaker ID information. We then excluded every line spoken by the interviewer, and every line in which the participant did not describe the picture. Such lines could be questions about the task (e.g., "Am I going too fast?"), reactions to the task (e.g., "Oh great."), or the participant declaring that he or she finished the task (e.g., "That's all I can see."). The reason for this exclusion is that such statements tend to be more formulaic. Because participants differed in the degree to which they made them, we regarded them as a potential confound. We then deleted any word repetitions which were likely the result of word-form retrieval or phonological problems. While these problems are an important aspect of clinical language research, repetitions of this sort do not reflect sentence representations which were our main concern.

Word combinations that crossed main clause boundaries were also excluded, unless they were connected lexically (e.g., "and", "or", "that"). For example, in the utterance "The mother's drawing the dishes, frowning but not turning off the faucet" the combinations "dishes, frowning", "the dishes, frowning" and "dishes, frowning but" were excluded. We also excluded word combinations that spanned across false starts. For example, in analyzing the utterance "The boy's in the cookie jar giving his … going in the cookie jar" we excluded "his … going", "giving his … going" and "his … going in". We further excluded word combinations which were ungrammatical. In most cases, these were the result of false starts and self-corrections. Word combinations were excluded by manually inserting a tag (a

period in angle brackets) where we saw a crucial boundary. The FLAT notes these "separator tags" and treats them like sentence final punctuation, meaning that word combinations crossing this boundary are not considered in the analysis.

The following is a transcript formatted for FLAT analysis. It is from a speaker with probable AD (264-1v-0). There are two syntactic errors ("reaching up in the cupboard", missing verb in "the boy on the stand or on that stool and…"), two semantic errors ("on the stand", "it rolled over on the floor") and ten weak inflections:

C:      kid's climbing up on the stool and reaching up in the cupboard .

C:      they aren't going to knock things off .

C:      and the mother <.> oh boy <.> the water's all spilling out of the sink .

C:      she's just looking at it like .

C:      oh <.> for goodness sakes .

C:      well as I say that <.> you know <.> the boy on the stand or on that stool and it's tilting .

C:      that's a good way to break his neck .

C:      break his back <.> I should have said .

C:      woman left her faucet running and it rolled over the floor .

### 2.2.2    Language analysis

The FLAT provides word, bigram and trigram counts. To measure connectivity in speech based on FLAT values, we divided the number of trigrams by the number of words. Higher outcomes indicate that more words are part of word combinations, therefore indicating more connected speech. Using an automated script written by Zimmerer in R [54] we further determined the rate of closed-class words within each sample. The script counts each word which matches a list comprising all English determiners, pronouns, number words, auxiliary verbs and adverbs other than manner adverbs [55]. Finally, we manually counted the number of weakly inflected words (verb inflections, plural inflections) as a measure of grammatical affixation, and semantic as well as grammatical errors to

calculate respective rates. If an error was considered part of the dialect (e.g., lack of verb inflection for third person singular) it was not counted. A second rater checked 60% of the samples. Interrater class coefficients showed moderate agreement for both syntactic (.565) and semantic (.5) errors. In the first stage of our analysis we investigated these variables ("general production measurements") at between- and within-group levels.

In the second stage, we focused on frequency and frequency-based variables. Observed frequencies relate to the occurrence of a word or n-gram in the spoken BNC (~10 million words). The FLAT uses additive smoothing, adding one to each frequency value, to avoid division by zero when computing variables based on observed frequencies.

N-gram frequency effects may be driven by the frequency of individual words: low frequency words tend to appear in low frequency n-grams. To address this issue and make sure that frequency effects occur at the level of combinations and therefore address questions about formulaic production, other statistics make use of the "expected frequency". Expected frequency is the number of occurrences one would expect based on the frequency of individual words if word order in the corpus was random. For a bigram *ab* and a trigram *abc* expected frequency is calculated as (with the approximate number of words in the spoken BNC 10,000,000):

$$expected\ frequency_{ab} = \frac{frequency_a\ frequency_b}{10,000,000}$$

$$expected\ frequency_{abc} = \frac{frequency_a\ frequency_b\ frequency_c}{10,000,000^2}$$

The difference between expected and observed frequencies suggests how much individual words are drawn to each other not by their individual frequencies, but by the patterns of normal language. It is indicative of formulaicity. We determined this difference using t-scores, which is one of the most commonly used collocation variables [56]. We selected t-scores over, for example, Mutual Information

since an analysis of all word combinations in our samples results in a high number of low frequency

units, and t-scores are less likely to be confounded by the resulting distribution [57]. A t-score for a

given n-gram *x* was calculated as:

$$t_x = \frac{frequency_x - expected\ frequency_x}{\sqrt{frequency_x}}$$

Table 1. Statistical values extracted using the Frequency in Language Analysis Tool (FLAT). Example values are provided for the sentence "I don't know", which the BNC segments into four words (and consequently, three bigrams and two trigrams).

| | I | do | n't | know |
|---|---|---|---|---|
| Word (observed) frequency | 30956 | 9956 | 12628 | 5743 |
| Bigram (observed) frequency | - | 2157 | 4222 | 1065 |
| Bigram t-score | - | 125.9 | 199.3 | 96 |
| Trigram (observed) frequency | - | - | 1866 | 904 |
| Trigram t-score | - | - | 136.3 | 95 |

In the third stage, we employed logistic and linear regression models to investigate the contribution of adding frequency-based variables to a model of language in AD. Table 1 contains FLAT values for the sentence *I don't know*. Box 1 provides a summary of all variables used in this study.

**Box 1. Overview of language variables**

**Number of words:** A measure of verbal responsiveness. Word counts exclude immediate repetitions as the result of production failure. Since the FLAT interfaces with BNC data structure, we applied the word segmentation from the BNC, meaning that contractions such as *don't* were counted as two words. Plural suffixes were also counted as individual words. Word count (and all other measures) included only utterances which were produced as part of the picture descriptions.

**Connectivity:** A measure of connected language, the proportion of words which appeared in grammatical trigrams. Higher values indicate that language was more connected, i.e., that there were fewer utterances one or two words in length. Connectivity was calculated as follows:

(number of trigrams / number of words) x 100%

**Rate of closed-class words:** Closed-class words (determiners, pronouns, number words, auxiliary verbs and adverbs other than manner adverbs) were automatically counted. Closed-class words are closely integrated with grammatical processing and may be easier to retrieve for speakers with lexical impairment. Rate of closed-class words was calculated as:

(number of closed-class words / number of words) x 100%

**Inflection rate:** Number of inflected words (verb inflections, plural inflections) was counted manually. Inflection rate was calculated as:

(number of inflected words / number of words) x 100%

**Syntactic error rate:** Syntactic errors were errors of inflection, agreement, argument structure or word order regardless of semantic content. Syntactic error rate was calculated as:

(number of syntactic errors / number of words) x 100%

**Semantic error rate:** Semantic errors represented incongruent lexical items resulting in descriptions that did not match the picture material. Semantic error rate was calculated as:

(number of semantic errors / number of words) x 100%

**Average word frequency:** Average frequency with which words appear in the BNC, in number of occurrences per million.

**Average bigram/trigram frequency:** Average frequency with which bigrams/trigrams appear in the BNC, in number of occurrences per million. Only n-grams which occur in the BNC were included.

**Average bigram/trigram t-score:** Average difference between the number of expected occurrences of the n-gram on the basis of frequency of its individual words, and the actual frequency with which it appears in the BNC (see 2.2.2 for formulas). Positive values indicate that n-gram appears more often than the expected frequency, and increasing distance from zero indicates increasing difference. Only n-grams which occur in the BNC were included.

**BiBNC/TriBNC rate:** Proportion of bigrams/trigrams which occur in the BNC, calculated as:

(number of BiBNC / number of bigrams) x 100%; (number of TriBNC / number of trigrams)

3. Results

3.1 Distributions, multiple comparisons and one- vs. two-tailed testing

Shapiro-Wilk tests were used for each sample to determine whether distributions were parametric. Dependent on the outcome, Independent Sample t-tests or Mann-Whitney U tests were applied for group comparisons. For effect size, we calculated r for comparisons between non-parametric samples. For parametric samples we provide Cohen's *d*, but also convert the value to r to allow better comparison. For within-group relationships between language and other outcomes Pearson's r was calculated for parametric and Kendall's tau rank for non-parametric samples.

In cases in which multiple comparisons were used to test one hypothesis, we provide Bonferroni corrected significance thresholds. However, t-scores are calculated using frequency, and the two variables are therefore strongly related (please see Appendix A for correlations). In this case Bonferroni corrections should be interpreted with caution since they may represent too strict a criterion and lead to Type II errors.

Based on our review of the literature and proposals of strong associations between AD and an increase in formulaicity, we predicted a negative correlation between co-occurrence variables and MMSE scores, and a positive correlation between co-occurrence variables and time post-onset. For this reason we report one-tailed tests (p values need to be doubled to determine the two-tailed statistic).

3.2 General production measurements (word count, connectivity, closed-class words, inflections, errors)

3.2.1 Between-group comparisons

The mean word count was 95.2 (SD = 39.8) for controls and 90.6 (SD = 39.6) for people with AD diagnosis. The difference was not significant, U = 971.5, p = .605; r = .06. The proportion of words

appearing in trigrams was 75% (SD = .07) for controls and 69% (SD = .07) for the AD group. That difference was significant, U = 1376, p < .001; r = .44.

The mean rate of closed-class words was 60% (SD = 5.9) for controls and 64% (SD = 6.6) for the AD group. The difference was significant, U = 612.5, p = .009; r = -.28. The inflection rate was 15.1% (SD = 4.4) for controls and 14.4% (SD = 10.7) for the AD group. The difference was at the threshold for statistical significance, U = 1137, p = .05; r = .21. The syntactic error rate (n of errors / words) was .009 for controls (SD = .0231) and .0141 (SD = .0179) for the AD group. The difference was significant, U = 636, p = .012; r = -.27. Semantic error proportion was .003 for controls (SD = .0233) and .005 (SD = .0081) for the AD group. The difference was significant, U = 652.5, p = .001; r = -.35.

After residualization over age and education, we found the following differences to be significant at p < .05: semantic error rate, t(65.464) = 2.443, p = 0.17; *d* = .94, r = .43, proportion of words within bigrams, t(84) = -3.258, p = .002, *d* = .64, r = .3, proportion of words within trigrams, t(84) = -2.893, p = .005. The difference in rate of closed-class words approached significance, t(83.996) = 1.87, p = .065; *d* = .46, r = .22.

### 3.2.2   Within-group analysis: Controls

We next looked at relationships between general production measurements in controls, starting with the non-linguistic measures of age, MMSE score and years of formal education. There was a significant relationship between age and MMSE scores, as older speakers tended to have lower scores, τ = -.269, p = .039. A positive correlation between MMSE scores and years of formal education showed a trend towards significance, τ = .232, p = .091.

We then investigated word count, the rate of closed-class words, the inflection rate, semantic errors, syntactic errors, and the proportion of words in trigrams as a measure of connected speech. An increase in number of words produced was associated with a higher rate of closed-class words, τ = .426, p < .001, a smaller inflection rate, τ = -.51, p < .001, and more connected speech, τ = .319,

p = .005. Speakers with more connected speech produced more closed-class words, τ = .367, p = .001, and fewer inflections, τ = -.341, p=.003. Speakers who produced a higher rate of closed-class words also had a lower inflection rate, τ = -.324, p = .004. Neither the proportion of syntactic nor of semantic errors correlated with other measures.

Only the number of words produced correlated with non-linguistic scores. Participants who produced more words tended to have higher MMSE scores, τ = .262, p = .043, and more years of formal education, τ = .252, p = .038.

### 3.2.3    Within-group analysis: People with probable AD

We followed the same analysis procedure as with the control group. However, we replaced age by the clinically more relevant variables onset age and time post-onset. Participants with more years of formal education tended to have a lower onset age, τ = -.246, p = .022. MMSE scores correlated with none of these values.

We then correlated the numbers of words produced, proportions of closed-class words and inflected words, and degree of connected speech. Speakers who produced more words produced more connected speech, τ = .294, p = .003, a higher rate of closed-class words, τ = .369, p < .001 and a smaller inflection rate, τ = -.36, p < .001. Rate of closed-class words was negatively correlated with inflection rate, τ = -.224, p = .025. Speakers who produced proportionally more syntactic errors produced more semantic errors, τ = .29, p = .012 and less connected speech, τ = -.21, p = .044. Of all these variables, only the syntactic error rate correlated with MMSE scores, as speakers with lower scores produced more errors, τ = -.256, p = .008. Syntactic errors also correlated positively with onset age, τ = .278, p = .008, and negatively with education, τ = -.238, p = .045.

### 3.2.4    Interim discussion

General production measurements can distinguish AD and control groups. Starting with the strongest variable, people with AD have less connected speech, make more semantic errors, use more closed-

class words, make more syntactic errors and produce fewer inflected word forms. The data show strong correlations between variables. Some of these relationships are conceptual: Since verbs and nouns are usually modified by inflections, the negative correlation between inflection and rate of closed-class words is expected. Similarly, higher verbal responsiveness is associated with increased connectivity in speech. In controls, poorer cognitive capacity as measured by the MMSE was associated with a lower verbal responsiveness. In the AD group, poorer cognitive capacity was associated with a higher number of syntactic errors. Speakers with later symptom-onset made more syntactic errors. This may suggest that syntactic errors are a marker for onset age. However, onset age was also negatively correlated with education, which is consistent with findings that suggest earlier onset and faster decline in people of higher educational status [58,59], but at odds with "cognitive reserve" hypotheses which predict later onset as the result of increased education [e.g., 60]. Strikingly, none of the general production measures correlated with time post-onset.

### 3.3 Frequency-based measurements

### 3.3.1 Floor effects in n-gram frequency distributions

We found that a proportion of bigrams and trigrams produced during Cookie Theft descriptions do not occur in the spoken BNC. The FLAT assigns these n-grams a frequency of 1 (due to additive smoothing, see above). This floor effect has consequences for the analysis. N-grams which do not appear in the spoken BNC can still differ substantially in how often they occur in natural language use, and our analysis would not be able to capture these differences. Further, t-scores as an indicator of formulaicity of word combinations are confounded if n-grams do not occur in the BNC, since in such cases they are solely determined by the frequency of individual words within the n-gram. Of two n-grams which do not occur in the reference corpus, the one that contains words with lower frequency has a higher t-score.

Of 7963 words analyzed in this study, 7940 occur in the spoken BNC (99.7%). Of the 6837 bigrams included, 6261 occur in the BNC (91.6%). Of 5807 trigrams included, 3400 occur in the BNC (58.6%).

As a consequence, we used a cutoff and included only n-grams which occur in the BNC. At the same time, we used the size of the floor effect in each individual as an additional indicator of formulaicity. Samples which contain more grammatical combinations that occur in the BNC can be considered more formulaic. We therefore calculated the proportion of bigrams and trigrams that occur in the BNC ("biBNC rate; "triBNC rate"). The variables capture the size of the floor effect for each individual sample. We associated a higher proportion of n-grams that occur in the BNC with a higher degree of formulaicity.

### 3.3.2 Between-group comparisons

Mean word frequency was 10850 (SD = 1557) for controls and 11003 (SD = 1493) for the AD group. The difference was not significant, U = 930, p = .876; r = .02. Of all bigrams produced by controls, 90% (SD = 6) occur in the spoken BNC. For the AD group, the proportion was 92% (SD = 4). This difference approached significance, U = 715.5, p = .087, r = -.18. The average bigram frequency was 324 (SD = 89) for the control group and 375 (SD = 131) for the AD group. The difference was significant, t(82.22) = 2.144, p = .035; $d$ = .57, r = .27. Average bigram t-score was 22.01 (SD = 5.03) for controls and 25.92 (SD = 6.59) for the AD group. The difference was significant, t(84) = 3.02, p = .003; $d$ = .78, r = .36.

Of the trigrams produced by controls, 52% (SD = 9) occur in the spoken BNC. That proportion was 60% (SD = 10) for the AD group. The difference was significant, U = 517.5, p = .001, r = -.37. The average trigram frequency was 30 (SD = 32) for the control group, and 34 (SD = 32) for the AD group. The difference was not significant, U = 782, p = .258; r = -.12. The average trigram t-score was 7.76 for the control group (SD = 3.19) and 9.08 (SD = 3.6) for the AD group. The difference approached significance, U = 700.5, p = .066; r = -.2.

After residualizing over age and education, differences in proportions of biBNC, t(83.689) = 2.382, p = .02; $d$ = .6, r = .29, and proportions of triBNC, t(84) = 3.367, p = .001; $d$ = .79, r = .37 were significant.

The difference between mean bigram t-scores approached significance, t(84) = 1.934, p = .056; *d* = .44, r = .21.

### 3.3.3    Within-group analysis: Controls

We compared frequency-based variables, namely word frequency, n-gram frequencies and t-scores, with other language related variables. We assumed that comparisons with bigram and trigram variables test hypotheses regarding the relationship between co-occurrence values and other areas. Since we chose six variables for these comparisons, a Bonferroni adjusted significance threshold would be p=.008 (however, see above for our reservations because of the close relationship between variables).

- Controls who produced more words had higher co-occurrence values. Number of words correlated with bigram t-scores, τ = .290, p = .011, and triBNC rate, τ = .415, p < .001.

- Higher rate of closed-class words also correlated positively with co-occurrence values, namely trigram frequency, τ = 25, p = .028, trigram t-scores, τ = .286, p = .012, and triBNC rate, τ = .286, p = .012.

- Inflection rate correlated negatively with co-occurrence values, in particular bigram frequency, τ = -.309, p = .006, bigram t-scores, τ = -.255, p = .024, biBNC rate, τ = -.253, p = .026, and triBNC rate, τ = -.345, p = .002.

- Speakers whose speech was more connected had higher co-occurrence values. Connected speech correlated with triBNC rate, τ = .325, p = .004.

- Proportion of syntactic and semantic errors correlated with none of the co-occurrence values variables.

We finally correlated corpus-based variables with age, years of formal education and MMSE scores. For MMSE scores, previous research suggests a negative correlation with co-occurrence values, and

we therefore calculated one-tailed correlations. However, there were no strong or significant correlations in neurotypical controls.

### 3.3.4 Within-group analysis: People with probable AD

We started by examining the relationship between frequency-based variables and other language variables.

- Speakers who produced more words had higher co-occurrence values, in particular trigram frequency, τ = .238, p = .017, trigram t-scores, τ = .251, p = .012, and triBNC proportion, τ = .249, p = .013.

- Rate of closed-class words was also associated with higher co-occurrence values, namely bigram frequency, r = .541, p < .001, bigram t-scores, r = .579, p < .001, trigram frequency, τ = .388, p < .001, trigram t-scores, r = .589, p < .001, biBNC rate, r = .389, p = .006, and triBNC rate, τ = .357, p < .001.

- Inflection rate was lower in speakers with higher co-occurrence values. Inflected words correlated with bigram frequency, τ = -.219, p = .028, bigram t-scores, τ = -.209, p = .036, trigram frequency, τ = -.271, p = .007, trigram t-scores, τ = -.272, p = .006, and triBNC rate, τ = -.234, p = .019.

- Connected speech and proportion of errors did not correlate with co-occurrence values.

We then correlated frequency-based variables with years of formal education, onset age, MMSE scores and time post-onset. Note that based on the literature we hypothesized negative correlations between co-occurrence and MMSE measures, and positive correlations between co-occurrence values and time post-onset, and we therefore used one-tailed correlations for these comparisons. Years of formal education correlated with one variable, biBNC rate, τ = .244, p = .021. Onset age correlated with none of the variables. Speakers with lower MMSE scores displayed higher co-occurrence values. MMSE scores correlated with bigram frequency, τ = -.225, p = .014, and trigram frequency, τ = -.194,

p = .03. The relationship with trigram t-scores was at the significance threshold, τ = -.168, p = .051.

Time post-onset correlated with bigram frequency, r = .254, p = .04, and bigram t-scores, r = .384,

p = .003 (Figure 1.).

Figure 1. Correlation between bigram t-scores and estimated years post symptom-onset in people

diagnosed with AD. Control mean and SD displayed for comparison.
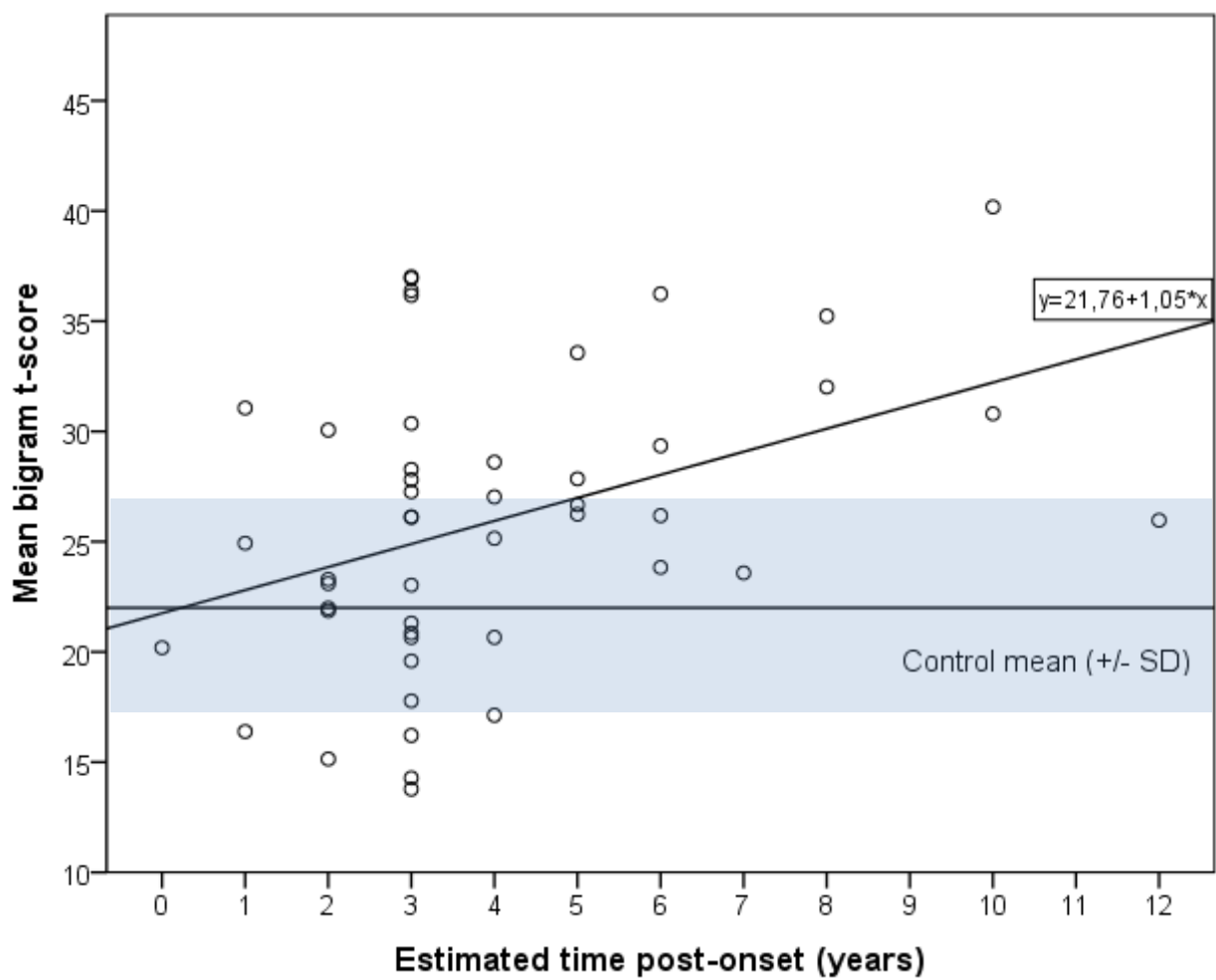
3.3.5    Interim discussion


Frequency-based variables could also distinguish between controls and AD groups. Compared to general production measurements, the predictive power of these variables was not as strong as that of speech connectivity, but similar to the rate of semantic errors. Higher co-occurrence values were associated with higher verbal responsiveness, a higher rate of closed-class words, a lower rate of inflected words, and only in the control group, more connected output. In the AD group, lower MMSE scores were associated with higher co-occurrence values. Of all language variables used in this study, co-occurrence variables were the only that correlated with time post-onset.


3.4  Models of language in AD

The final step in our analysis was to determine the degree to which language output, as elicited via Cookie Theft descriptions, could predict categorization into AD and control groups, as well as MMSE scores, onset age and time post-onset within the AD group. Given the high intercorrelational relationships between variables, we were particularly interested in whether frequency-based variables add predictive power if used together with the other variables. We chose all non-frequency-based variables (word count, connectivity, rate of closed-class words, inflection rate, proportion of semantic and syntactic errors) as a baseline for comparison against models which added the variables extracted from the BNC. For frequency-based variables, we selected word frequency, and for bi- and trigrams, t-scores over raw frequencies given the greater independence of t-scores from word frequency. We therefore generated five variable sets:

A.  Word count, Connectivity, Rate of closed-class words, Inflection rate, Semantic error rate and Syntactic error rate

B.  Set A, Word frequency, bigram t-scores and BiBNC rate

C. Set A, Word frequency, trigram t-scores and TriBNC rate

We ranked the values across both groups for between-group comparisons, and within the AD group for investigating MMSE, onset age and time post-onset. Educational status correlated with onset age. To avoid potential confound we residualized onset age over years of formal education and used the residuals in our models.

We used logistic regressions to see how the variable sets predict categorization into the two speaker groups. When all variables were entered, variables of Set A distinguished between people with probable AD and controls, $\chi^2$ = 45.896, p < .001, df = 6. Nagelkerke's $R^2$ of .554 indicated a moderately strong relationship between predictors and grouping. Three of the variables in the set contributed to the model at significant levels (starting with the strongest predictor): Connectivity, Wald = 12.337, p < .001, Rate of closed-class words, Wald = 9.329, p = .002, and Semantic error rate, Wald = 4.563, p = .033. A forward conditional model, which starts with the strongest predictor and only adds new predictors if they significantly improve the model, included only these three variables. This reduced model predicted grouping with very similar strength, $\chi^2$ = 42.713, p < .001, df = 3, Nagelkerke's $R^2$ = .524.

In Set B, frequency variables made no contribution, and conditional models selected the same three variables as in Set A. Set C however showed a small, but significant contribution of TriBNC rate. The full model showed a strong relationship between predictors and grouping, $\chi^2$ = 55.102, p < .001, df = 8, Nagelkerke's $R^2$ = .634. Two variables significantly contributed to the model: Connectivity, Wald = 11.661, p < .001, and TriBNC rate, Wald = 7.51, p = .006. Semantic error rate approached significance, Wald = 3.534, p = .06, as did Rate of closed-class words, Wald = 3.265, p = .071. A forward conditional model selected these four variables and showed a strong relationship between predictors and grouping, $\chi^2$ = 51.851, p < .001, df = 4, Nagelkerke's $R^2$ = .606.

None of the variable sets predicted MMSE at significant levels. In linear regression models, only Set C showed a trend, $F_{(8, 39)} = 1.828$, $p = .095$, $R^2 = .302$. Trigram t-scores made a contribution approaching significance, $\beta = -.032$, $p = .082$. All stepwise models selected a single variable, Syntactic error rate. The model made a weak, but significant prediction, $F_{(1, 46)} = 6.45$, $p = .015$, $R^2 = .123$.

Equally, Onset age (residualized over education) was predicted best by Syntactic error rate. Set A was the best combined model, but did not reach significance, $F_{(6, 41)} = 2.182$, $p = .064$, $R^2 = .242$. Syntactic error rate made the only significant contribution, $\beta = .404$, $p = .011$, while the contribution of Inflection rate approached significance, $\beta = .314$, $p = .063$. For all sets, stepwise models selected Syntactic error rate only, $F_{(1, 46)} = 5.282$, $p = .026$, $R^2 = .103$.

With regards to years post-onset, co-occurrence values variables made moderate and significant contributions when predicting estimated years post-onset for the AD group. Set A was weak, $F_{(6, 41)} = 1.445$, $p = .22$, $R^2 = .175$. The contribution of Rate of closed-class words approached significance, $\beta = .306$, $p = .078$. The stepwise model selected none of the variables. Set B significantly predicted time post-onset, $F_{(8, 39)} = 2.305$, $p = 0.35$, $R^2 = .353$. Bigram t-scores was the only variable which made a significant contribution, $\beta = .416$, $p = .021$. A stepwise model made significant predictions, $F_{(3, 44)} = 6.066$, $p = .002$, $R^2 = .293$, and selected three variables: Bigram t-scores, $\beta = .456$, $p = .001$, Syntactic error rate, $\beta = -.288$, $p = .03$, and BiBNC rate, $\beta = -.276$, $p = .036$. Trigram data were weaker. Set C was not significant, $F_{(8, 39)} = 1.352$, $p = .244$, $R^2 = .244$. Only the contribution of Trigram t-scores approached significance, $\beta = .331$, $p = .083$. For the latter two sets, stepwise regressions selected none of the variables.

Discussion

This study presents a first foray into investigating language in AD using variables based on frequency of use. Previous reports suggest that people with AD show increased use of formulas [40,44]. We

aimed to establish whether frequency and collocation strength of word combinations can help distinguish healthy speakers from speakers with a diagnosis of AD, and track disease progression in pathological groups. We applied the FLAT, an automated tool which determines frequency and related variables for word bigrams and trigrams by data mining the spoken portion of the BNC. It is an approach which is fast, blind, but solely focuses on frequency, not function. Using correlations and regression models, we explored the relationship between co-occurrence variables and other language variables and the contribution of adding the former to a language model of AD.

Our investigation of data from DementiaBank's Pitt Corpus, comprising Cookie Theft picture descriptions, showed that a range of usage-based variables can distinguish between AD and healthy language, after accounting for differences in age and educational status. These variables also correlated with MMSE scores in the patient group. Notably, only co-occurrence values were correlated with estimated time post-symptom onset. Word frequency, on the other hand, showed no predictive strength in our data set. It may be a limitation of the Cookie Theft elicitation task that it narrows lexical breadth by requiring description of one particular picture. Of the closely related frequency-based variables we tested, bigram t-scores in particular seemed to offer the most sensitivity. However, there is some evidence that determining the proportion of trigrams that were below our cutoff can add predictive power.

It would seem plausible that data from larger word combinations (e.g., tetragrams) would be more powerful. However, increasing n-gram length comes with a decrease in raw frequency, contributing to a floor effect in which a proportion of n-grams do not occur in the BNC at all. This results in a loss of information that becomes apparent when comparing bigram to trigram data in our study. Trigram data was generally a weaker predictor, likely because of the bigger floor effect.

In relation even to the most effective co-occurrence values variables, other language variables can be better predictors for some comparisons. In group distinctions, connectivity of output and semantic errors showed greater effect sizes. Lower connectivity can be related to difficulties in word retrieval,

resulting in false starts and unfinished utterances, or structural simplification of output. However, the biggest predictor for AD diagnosis was not a language variable, but MMSE scores. This is no surprise since MMSE or similar tests are a crucial diagnostic measure, and our study contrasted heterogeneous MMSE performances in the AD group with data from healthy controls who typically score at ceiling. However, MMSE performance showed no sensitivity to disease progression as measured by time post-onset.

In regression models distinguishing AD and controls, connectivity, closed-class words and semantic errors were clearly the best predictors. Trigram data made a smaller, but significant contribution by additionally explaining about 8% of the distribution. As predictors for MMSE, syntactic error rate was the only, but only weak, predictor. Similarly, syntactic errors were the sole, and weak, predictor for onset-age (with educational status accounted for). For time post-onset, bigram data were the best predictor. A model which contained them and syntactic error rate displayed a moderate relationship. A cautious conclusion on the basis of this first exploration would be that errors and connectivity are more clearly mapped to the large cognitive differences that distinguish the two groups, or the substantial changes that are associated with decreasing MMSE scores. Co-occurrence values variables, on the other hand, represent a more fine-grained scale with potential to track subtle language change, and may be best suited to track the often slow cognitive changes that occur during the prodromal period, with increasing time post-onset or through intervention.

To test the power of these variables, we allowed the models to directly compete with the other language variables. As stated in the introduction, this competition can be unjustified if one adopts holistic language models. Closed-class words in particular are often part of formulas or formulaic frames [52]. For instance, the sentence "I don't know" consists mostly of function words. The strong relationship between frequency-based variables and rate of closed-class words is likely a result of this relationship.

The FLAT is work in progress and we are exploring ways to improve it. The issue of longer n-grams being less effective may be resolved by either increasing the size of individual samples or of the reference corpus. Using the complete BNC with 100 million words instead of the spoken subcorpus with 10 million may be an improvement despite the resulting dominance of written text. However, this tenfold increase would overload computational resources, and analysis would become much slower. This would act against our aim of turning the FLAT into an easily used resource (although computational power will increase with time as technology improves).

FLAT analyses are solely based on frequency, and not on function. Including parts-of-speech information (i.e., grammatical classes) and sentence types may start to close the gap between our current work and traditional analyses, but ultimately, a full account of formulaicity in language has to include an analysis of meaning. At the same time we recommend excluding some variables in the future in order to simplify the model. Based on our data, inflection rate may be irrelevant at least for investigations of language in AD.

Our results strongly support the view that formulaicity increases as the result of AD. Classic generativist models [61,62] do not take usage variables such as frequency into account and therefore cannot integrate our results without postulating additional systems. Formulaicity fits to connectionist frameworks [63] which regard grammatical processing as connections between lexical "nodes" which can excite and inhibit each other. These models are very dynamic and change their weights with every exposure to language, strengthening connections if they are more frequently needed. Connectionist models predict that more strongly collocated combinations become easier to process, even if they are of the same structural category (e.g., transitive actives).

However, connectionist models are only weakly holistic since they commonly have, at their basis, a layer in which every morpheme is represented as a single unit. Another usage-based approach, construction grammar, has been proposed most prominently by Goldberg [64,65] and Tomasello [30]. Construction grammar predicts that word combinations can be completely lexicalized and treated like

a single word. This can be the result of early exposure during acquisition, or frequent use over the lifespan. Phonetic data, especially on contractions (e.g., "wanna", "gonna", "I dunno"), support this notion [11]. Lexicalized (or "fossilized") expressions may not involve any combinatorial processes. These two phenomena – overlearned combinations vs. complete lexicalizations – are fundamentally different, but not exclusive. It is possible that some formulas are the result of strengthened connections, and others are represented as single lexical unit. It is also possible that some formulas can be processed as both simultaneously: "I don't know" may be fully lexicalized, but our language system can still analyze its constituents. On this issue there is only little empirical work [see 66, for a review].

Similarly, the nature of formulaicity in AD remains to be explained. According to Ullman's [67,68] declarative/procedural model of language, stored elements such as words and functional morphemes fall under declarative memory, while word combinations (grammar) are procedural. AD is commonly described as a breakdown of declarative memory, and Bridges and van Lancker Sidtis [44] see formulas as procedural, habitual combinations. They liken the preservation of formulas in AD to the retained ability to play a musical instrument or card game. We note however that the declarative/procedural model of language was created to accommodate a words and rules approach. Combinatorial processes do not only require procedural memory, but also demand integration with declarative memory, since each word and morpheme needs to be retrieved from the mental lexicon.

A solution may be to focus on the declarative aspect of word combinations as well as on possible lexicalization. Ullman [68] speculates that phrases and sentences are stored in the lexicon. If more frequent combinations are more strongly anchored within declarative memory than less frequent combinations, impairment of declarative memory would result in production of more frequent utterances. At the same time, production of novel utterances would be constrained by difficulties accessing multiple words and functional morphemes. Generally, work on holistic language processing in other areas such as language acquisition has challenged the words vs. rules dichotomy. Instead,

some see language as involving a lexicon-syntax continuum [1,64] that may ultimately challenge a strict binary notion of declarative vs. procedural processing for explaining pathologies.

Our research program sees considerable potential in automatic, or semi-automated language analysis in diagnosis and classification, and such analyses, using very different approaches, are being implemented in dementias [27,69], schizophrenia [70], and language acquisition [71]. Our study demonstrates how formulaic language can be analyzed quickly and blindly using frequency and frequency-based variables, and suggests that addition of these variables can aid identification and tracking of dementia even on the basis of small individual samples. Like other approaches to automatic analysis and classification, our results are in need of replication. In particular, we are interested in looking at longitudinal samples instead of making use of within-group differences in cross-sectional samples. Such validation is crucial especially given that the variables of estimated onset age and symptom onset may be confounded by individual's abilities to detect change in behavior, and their willingness to seek medical advice. While the Pitt Corpus contains follow-up measurements, drop-out rates are high and a follow-up based on our initial selection criteria lacks sufficient power. Groups could be expanded by adding participants with additional diagnoses (such as depression). Such a step would approach the complexity of clinical reality, but also complicate statistical models beyond what we considered useful for the current study. However, on the basis of our results we believe that with increasing understanding of formulaicity variables, research into language formulas can make a substantial contribution to diagnosis and monitoring of dementia.

References:

[1]     Croft W (2001) *Radical construction grammar: Syntactic theory in typological perspective*, Oxford University Press, Oxford.

[2]     Chomsky N (1957) *Syntactic Structures*, Mouton, The Hague.

[3]     Chomsky N (1965) *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

[4]     Pinker S (1999) *Words and rules: The ingredients of language*, Basic Books, New York, NY.

[5]     Wray A (2012) What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annu. Rev. Appl. Linguist.* **32**, 231–254.

[6]     Wray A, Perkins MR (2000) The functions of formulaic language: an integrated model. *Lang. Commun.* **20**, 1–28.

[7]     Conklin K, Schmitt N (2012) The processing of formulaic language. *Annu. Rev. Appl. Linguist.* **32**, 45–61.

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

[9]     Sosa A, MacFarlane J (2002) Evidence for frequency-based constituents in the mental lexicon: collocations involving the word of. *Brain Lang.* **83**, 227–236.

[10]    Beckner C, Bybee J (2009) A usage-based acount of constituency and reanalysis. *Lang. Learn.*

**59**, 29–48.

[11]    Bybee J (2006) From usage to grammar: The mind's response to repetition. *Language (Baltim).* **82**, 711–733.

[12]    Scheibman J (2000) I dunno: A usage-based account of the phonological reduction of don't in American English conversation. *J. Pragmat.* **32**, 105–124.

[13]    Wray A (2014) Dementia and language. In *The Encyclopedia of Applied Linguistics*, Chapelle CA, ed. Wiley-Blackwell, Oxford.

[14]    Ripich DN, Terrell BY (1988) Patterns of discourse cohesion and coherence in Alzheimer's Disease. *J. Speech Hear. Disord.* **53**, 8.

[15]    Rousseaux M, Sève A, Vallet M, Pasquier F, Mackowiak-Cordoliani MA (2010) An analysis of communication in conversation in patients with dementia. *Neuropsychologia* **48**, 3884–90.

[16]    Chapman SB, Ulatowska HK, King K, Johnson JK, McIntire DD (1995) Discourse in early Alzheimer's Disease versus normal advanced aging. *Am. J. Speech-Language Pathol.* **4**, 124.

[17]    Badecker W, Caramazza A (1985) On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology: The case against agrammatism. *Cognition* **20**, 97–125.

[18]    Gola KA, Thorne A, Veldhuisen LD, Felix CM, Hankinson S, Pham J, Shany-Ur T, Schauer GP, Stanley CM, Glenn S, Miller BL, Rankin KP (2015) Neural substrates of spontaneous narrative production in focal neurodegenerative disease. *Neuropsychologia* **79**, 158–171.

[19]    Pekkala S, Wiener D, Himali JJ, Beiser AS, Obler LK, Liu Y, McKee A, Auerbach S, Seshadri S, Wolf PA, Au R (2013) Lexical retrieval in discourse: an early indicator of Alzheimer's dementia. *Clin. Linguist. Phon.* **27**, 905–21.

[20]    Mickes L, Wixted JT, Fennema-Notestine C, Galasko D, Bondi MW, Thal LJ, Salmon DP (2007)

Progressive impairment on neuropsychological tasks in a longitudinal study of preclinical Alzheimer's disease. *Neuropsychology* **21**, 696–705.

[21] Weiner MF, Neubecker KE, Bret ME, Hynan LS (2008) Language in Alzheimer's disease. *J. Clin. Psychiatry* **69**, 1223–7.

[22] Cuetos F, Rodríguez-Ferreiro J, Sage K, Ellis AW (2012) A fresh look at the predictors of naming accuracy and errors in Alzheimer's disease. *J. Neuropsychol.* **6**, 242–56.

[23] Cuetos F, Arce N, Martínez C, Ellis AW (2015) Word recognition in Alzheimer's disease: Effects of semantic degeneration. *J. Neuropsychol.*

[24] Sajjadi SA, Patterson K, Tomek M, Nestor PJ (2012) Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology* **26**, 847–866.

[25] Kempler D, Curtiss S, Jackson C (1987) Syntactic preservation in Alzheimer's disease. *J. Speech Lang. Hear. Res.* **30**, 343.

[26] Bates E, Harris C, Marchman V, Wulfeck B, Kritchevsky M (1995) Production of complex syntax in normal ageing and Alzheimer's disease. *Lang. Cogn. Process.* **10**, 487–539.

[27] Fraser KC, Meltzer JA, Rudzicz F (2015) Linguistic Features identify Alzheimer's disease in narrative speech. *J. Alzheimers. Dis.* **49**, 407–422.

[28] Kemper S, Marquis J, Thompson M (2001) Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychol. Aging* **16**, 600–14.

[29] Ahmed S, Haigh A-MF, de Jager CA, Garrard P (2013) Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* **136**, 3727–3737.

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

[31]    Lieven E, Salomo D, Tomasello M (2009) r. *Cogn. Linguist.*

[32]    Janssen N, Barber HA (2012) Phrase frequency effects in language production. *PLoS One* **7**, e33202.

[33]    Conklin K, Schmitt N (2008) Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Appl. Linguist.* **29**, 72–89.

[34]    Arnon I, Snider N (2010) More than words: Frequency effects for multi-word phrases. *J. Mem. Lang.* **62**, 67–82.

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

[36]    Code C (1982) Neurolinguistic analysis of recurrent utterance in aphasia. *Cortex* **18**, 141–152.

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

[38]    Zimmerer VC, Cowell PE, Varley RA (2014) Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia* **53**, 25–38.

[39]    Bayles KA (2004) Language: Dementia. In *The MIT encyclopedia of communication disorders*, Kent R, ed. MIT Press, Cambridge, MA, pp. 297–322.

[40]    Wray A (2011) Formulaic language as a barrier to effective communication with people with Alzheimer's disease. *Can. Mod. Lang. Rev.* **67**, 429–458.

[41]    van Lancker-Sidtis D, Rallon G (2004) Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Lang. Commun.* **24**, 207–240.

[42]    van Lancker Sidtis D, Postman WA (2006) Formulaic expressions in spontaneous speech of left- and right-hemisphere-damaged subjects. *Aphasiology* **20**, 411–426.

[43]    Sidtis D, Canterucci G, Katsnelson D (2009) Effects of neurological damage on production of formulaic language. *Clin. Linguist. Phon.* **23**, 270–284.

[44]    Bridges KA, Van Lancker Sidtis D (2013) Formulaic language in Alzheimer's disease. *Aphasiology* **27**, 799–810.

[45]    van Lancker Sidtis D, Choi J, Alken A, Sidtis JJ (2015) Formulaic language in Parkinson's disease and Alzheimer's disease: omplementary effects of subcortical and cortical dysfunction. *J. Speech. Lang. Hear. Res.* **58**, 1493–507.

[46]    Lieven E, Behrens H, Speares J, Tomasello M (2003) Early syntactic creativity: a usage-based approach. *J. Child Lang.* **30**, 333–370.

[47]    Dąbrowska E (2014) Recycling utterances: A speaker's guide to sentence processing. *Cogn. Linguist.* **25**, 617–653.

[48]    Zimmerer VC, Wibrow M (2015) Frequency in language analysis tool.

[49]    MacWhinney B (2007) TalkBank Project. In *Creating and Digitizing Language Corpora: Synchronic Databases, Vol. 1*, Beal JC, Corrigan KP, Moisl HL, eds. Palgrave-Macmillan, Houndmills.

[50]    Goodglass H, Kaplan E (1983) *The assessment of aphasia and related disorders*, Lea and Febiger, Philadelphia.

[51]    Folstein M, Folstein S, McHugh P (1975) "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **3**, 189–98.

[52]    Biber D (2009) A corpus-driven approach to formulaic language in English. *Int. J. corpus Linguist.* **14**, 275–311.

[53]    Quade D (1967) Rank analysis of covariance. *J. Am. Stat. Assoc.* **62**, 1187–1200.

[54]    R Development Core Team (2008) R: A language and environment for statistical computing.

[55]    Function words in English, Last updated 2014, Accessed in 2015.

[56]    Gries ST (2010) Useful statistics for corpus linguistics. In *A mosaic of corpus linguistics:*

*selected approaches*, Sánchez A, Almela M, eds. Peter Lang, Frankfurt am Main, pp. 269–291.

[57]     Evert S (2009) Corpora and collocations. In *Corpus Linguistics: An International Handbook Vol. 2*, Lüdeling A, Kytö M, eds. , Berlin, New York, pp. 1212–1248.

[58]     Scarmeas N, Albert SM, Manly JJ, Stern Y (2006) Education and rates of cognitive decline in incident Alzheimer's disease. *J. Neurol. Neurosurg. Psychiatry* **77**, 308–16.

[59]     Del Ser T (1999) An autopsy-verified study of the effect of education on degenerative dementia. *Brain* **122**, 2309–2319.

[60]     Willis KJ, Hakim AM (2013) Stroke prevention and cognitive reserve: emerging approaches to modifying risk and delaying onset of dementia. *Front. Neurol.* **4**, 13.

[61]     Chomsky N (1981) *Lectures on Government and Binding*, Foris, Dordrecht.

[62]     Chomsky N (1995) *Minimalist program*, MIT Press, Cambridge, MA.

[63]     Elman J, Karmiloff-Smith A, Bates E, Johnson M (1996) *Rethinking Innateness: A Connectionist Perspective on Development*, MIT Press, Cambridge, MA.

[64]     Goldberg AE (2003) Constructions: a new theoretical approach to language. *TRENDS Cogn. Sci.* **7**, 219–224.

[65]     Goldberg AE (2006) *Constructions at work: The nature of generalization in language*, Oxford University Press, New York.

[66]     Siyanova-Chanturia A (2015) On the "holistic" nature of formulaic language. *Corpus Linguist. Linguist. Theory* **11**, 285–301.

[67]     Ullman MT (2004) Contributions of memory circuits to language: the declarative/procedural model. *Cognition* **92**, 231–270.

[68]     Ullman MT (2001) A neurocognitive perspective on language: the declarative/procedural model. *Nat. Rev. Neurosci.* **2**, 717–26.

[69]     Fraser KC, Meltzer JA, Graham NL, Leonard C, Hirst G, Black SE, Rochon E (2012) Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*.

[70]     Elvevåg B, Foltz PW, Rosenstein M, Delisi LE (2010) An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J. Neurolinguistics* **23**, 270–284.

[71]     Barcelo-Coblijn L, Corominas-Murtra B, Gomila A (2012) Syntactic trees and small-world networks: syntactic development as a dynamical process. *Adapt. Behav.* **20**, 427–442.

Appendix A. Correlations between BNC-extracted frequency variables.

| | Bigram Frequency | biBNC frequency | Bigram t-score | biBNC t-score | Trigram frequency | triBNC frequency | Trigram t-score | triBNC t-score | biBNC proportion | triBNC proportion |
|---|---|---|---|---|---|---|---|---|---|---|
| Word frequency | τ = .224* p = .002 | τ = .168* p = .022 | τ = .075 p = .307 | τ = .024 p = .743 | τ = .061 p = .406 | τ = .05 p = .5 | τ = .087 p = .239 | τ = .07 p = .338 | τ = .227* p = .002 | τ = .127 p = .083 |
| Bigram frequency | | τ = .886** p < .001 | τ = .698** p < .001 | τ = .661** p < .001 | τ = .509** p < .001 | τ = .446** p < .001 | τ = .544** p < .001 | τ = .481** p < .001 | τ = .175* p = .017 | τ = .395** p < .001 |
| biBNC frequency | | | τ = .658** p < .001 | τ = .688** p < .001 | τ = .482 p < .001 | τ = .428** p < .001 | τ = .511** p < .001 | τ = .463** p < .001 | τ = .061 p = .41 | τ = .321** p < .001 |
| Bigram t-score | | | | τ = .878** p < .001 | τ = .524 p < .001 | τ = .449** p < .001 | τ = .602* p < .001 | τ = .515** p < .001 | τ = .162* p = .028 | τ = .503** p < .001 |
| biBNC t-score | | | | | τ = .487** p < .001 | τ = .429** p < .001 | τ = .559** p < .001 | τ = 494** p < .001 | τ = .041 p = .581 | τ = .409** p < .001 |
| Trigram frequency | | | | | | τ = .893** p < .001 | τ = .836** p < .001 | τ = .858 p < .001 | τ = .149* p = .042 | τ = .348** p < .001 |
| triBNC frequency | | | | | | | τ = .747** p < .001 | τ = .858** p < .001 | τ = .149* p = .042 | τ = .348** p < .001 |

| | | | | | | | | $\tau = .841^{**}$ | $\tau = .173^{*}$ | $\tau = .451^{**}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Trigram t-score | | | | | | | | $p < .001$ | $p = .019$ | $p < .001$ |
| triBNC t-score | | | | | | | | | $\tau = .1$ $p = .173$ | $\tau = .291^{**}$ $p < .001$ |
| biBNC proportion | | | | | | | | | | $\tau = .401^{**}$ $p < .001$ |