University of Groningen

Quantifying the Uncertainty of Parameters Measured in Spontaneous Speech of Speakers With Dementia

Ossewaarde, Roelant; Jonkers, Roel; Jalvingh, Fedor; Bastiaanse, Roelien

Published in:
Journal of speech, language, and hearing research

DOI:
[10.1044/2020_JSLHR-19-00222](https://doi.org/10.1044/2020_JSLHR-19-00222)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Ossewaarde, R., Jonkers, R., Jalvingh, F., & Bastiaanse, R. (2020). Quantifying the Uncertainty of Parameters Measured in Spontaneous Speech of Speakers With Dementia. *Journal of speech, language, and hearing research*, 63(7), 2255-2270. https://doi.org/10.1044/2020_JSLHR-19-00222

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Research Article

Quantifying the Uncertainty of Parameters Measured in Spontaneous Speech of Speakers With Dementia

Roelant Ossewaarde,^{a,b} Roel Jonkers,^a Fedor Jalvingh,^a and Roelien Bastiaanse^{a,c}

Purpose: Corpus analyses of spontaneous language fragments of varying length provide useful insights in the language change caused by brain damage, such as caused by some forms of dementia. Sample size is an important experimental parameter to consider when designing spontaneous language analyses studies. Sample length influences the confidence levels of analyses. Machine learning approaches often favor to use as much language as available, whereas language evaluation in a clinical setting is often based on truncated samples to minimize annotation labor and to limit any discomfort for participants. This article investigates, using Bayesian estimation of machine learned models, what the ideal text length should be to minimize model uncertainty.

Method: We use the Stanford parser to extract linguistic variables and train a statistic model to distinguish samples by

speakers with no brain damage from samples by speakers with probable Alzheimer's disease. We compare the results to previously published models that used CLAN for linguistic analysis.

Results: The uncertainty around six individual variables and its relation to sample length are reported. The same model with linguistic variables that is used in all three experiments can predict group membership better than a model without them. One variable (concept density) is more informative when measured using the Stanford tools than when measured using CLAN.

Conclusion: For our corpus of German speech, the optimal sample length is found to be around 700 words long. Longer samples do not provide more information.

The analysis of spontaneous language is important for corpus researchers interested in how ordinary language and its use differ between groups (Prins & Bastiaanse, 2004). One use case for spontaneous speech analysis, and the language it expresses is to measure the influence of dementia on language production. That influence can be quantified using statistical methods applied to a corpus of spontaneous language. A trained statistical model can predict group membership (speakers with brain damage vs. speakers with no brain damage [NBD]) based on speech transcripts (Fraser et al., 2014, 2013; Orimaye & Golden, 2014; Pakhomov et al., 2010;

Peintner et al., 2008; Roark et al., 2011; Szatloczki et al., 2015).

Such approaches require training data that are sufficiently abundant for a model to reliably learn to differentiate between groups. Abundance of data is a function of two aspects: enough training instances (conversations) and enough distributional difference between the variables of the model. It is essential to collect the right amount of data: enough for distinctions in the data to become meaningful, but not so much that the data collection becomes too extensive or places an undue burden on participants. It is an ethical obligation in all research to limit the burden on participants as much as possible. Some people with communication disorders, such as aphasia, struggle with protocols that require them to talk for an extended time. When the study population includes participants from such groups, that ethical obligation can be partially met by determining reasonable bounds for the amount of data to be collected—enough data so that an effect is detected if there is an effect, but not more than necessary.

The length of a speech sample is a concern as well for clinicians who evaluate oral or written language as part of their diagnostic testing. Too much text may be

^aCenter for Language and Cognition Groningen, Rijksuniversiteit Groningen, the Netherlands

^bInstitute for ICT, HU University of Applied Science, Utrecht, the Netherlands

^cCenter for Language and Brain, NRU Higher School of Economics, Moscow, Russia

Correspondence to Roelant Ossewaarde: r.a.ossewaarde@rug.nl

Editor-in-Chief: Sean M. Redmond

Editor: Sarah Elizabeth Wallace

Received September 18, 2019

Revision received December 8, 2019

Accepted February 25, 2020

https://doi.org/10.1044/2020_JSLHR-19-00222

Disclosure: The authors have declared that no competing interests existed at the time of publication.

burdensome for a patient to produce and is costly to analyze; too little may undercut the certainty of the diagnosis.

There are different ways to elicit data. Semispontaneous speech tasks involve pictures or cartoons that impose some sort of structure to the output. Fragment duration is typically 2–4 min (Boschi et al., 2017). Spontaneous speech tasks are often interviews and conversations with minimally predetermined structure or content, except for the questions that are asked. Spontaneous speech is sometimes longer than semispontaneous speech because there are fewer limitations inherent to the task of describing a given scene or story (Boschi et al., 2017; Prins & Bastiaanse, 2004).

Corpus linguists study linguistic phenomena based on the evaluation of some kind of frequencies, using corpora as their primary data (Gries, 2009). Clinical linguists sometimes use a frequency analysis for diagnostic purposes but generally use one or a few fragments from a single source as primary data. In clinical practice, linguists often truncate fragments for assessment, whereas corpus researchers hardly ever do this when they train statistical models through machine learning. Truncation is sometimes used if models include variables that depend on sample size in a nonlinear way, such as the type–token ratio (Malvern et al., 2004; Montag et al., 2018). Truncation also makes analysis more practical, and it is safe under the assumption that the information gain from further analysis decreases beyond a set point. For example, Brookshire and Nicholas (1995) and Vermeulen et al. (1989) base their analyses on the first 300 words of a fragment. Some clinicians use a set amount of time, rather than words, if the speech rate is low or severely nonfluent. It is an open question how much influence text length has on the performance of machine learned models that are used to classify language fragments into diagnostic groups.

Quantifications of language involve the coding of variables that capture the properties of the text. Variables may be calculated at different levels, for example, at the lexical level (e.g., age of acquisition of a word), clausal level (e.g., number of verbs in a clause), or the sentence level (e.g., number of relative clauses). In a given text, the number of words is almost by definition higher than the number of clauses or sentences. If fewer units are available for analysis, the random error may be larger, leading to more inflation of p values in statistical tests for higher level variables than for lower level variables if text length remains constant. If a specific degree of statistical stability is sought, the truncation limits of data may have to vary with the level of abstraction at which a variable is calculated.

In clinical practice, variables that are easy to observe are typically used for the assessment of spontaneous speech (Wilson et al., 2018). The use of software-based language analyses makes it easier to include sophisticated measures, often higher level variables, such as parse tree depth (Yngve, 1961) or the degree of formulaic language (Zimmerer et al., 2016) in predictive models. Such predictive models, typically derived using machine learning techniques, have been successful in detecting specific patterns associated with inter alia formal thought disorder (Çokal et al., 2018), Huntington's disease (Hinzen et al., 2018), and autism (Nakai et al.,

2017). Once such models mature enough, they can be used in a clinical context; for example, the Aachener Sprachanalyse (Husmann et al., 2012) is a computer-assisted method for the quantitative analysis of German spontaneous speech to assess aphasia. It uses distributional differences between aphasic and unimpaired language and is sensitive enough to detect changes over time in samples of persons with aphasia (Grande et al., 2008).

Software-based language analysis makes it more cost-effective to compute and include higher level variables. Because higher level variables may need longer texts for their computation, different models may pose different requirements on the length of their input texts. When the mix of lower and higher level variables changes, text length requirements may need to be reevaluated. For this reevaluation, a principled method is sought so that researchers can determine how much text material is needed to reliably measure variables of each kind.

This is a case study of how the minimally required text length for a given corpus can be determined so that models can be construed, which can make meaningful discriminations based on the predictor parameters. It is inspired by the questions that arose while determining parameters, in particular, the required text length, for a future longitudinal study of the decline of language in Dutch-speaking persons with dementia. The aim of that study is to follow groups of patients with primary progressive aphasia and map the decline of their language production over time.

The research question of this study is “How does the uncertainty around the estimation of model parameters in a predictive model vary as a function of the length of texts used for model training?” As training data, we use a convenience sample from a corpus of standard German spontaneous speech partly of neurodegenerated speakers. That corpus was compiled in the context of an earlier study on verbs and time reference in German. A desirable property of the corpus is that fragments are relatively long, so that it is possible to experiment with high truncation limits.

As predictive model, we use a model that was used in two prior studies that report significant results using regression models on labeled language data of English speakers. Orimaye and Golden (2014) used the Pitt data set from DementiaBank to analyze English language. The data used are semispontaneous narratives, a description of the Cookie Theft picture of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). de Lira et al. (2011) use a similar design, based on the “Dog Story” picture sequence. Both studies build a diagnostical model using variables that describe properties of the language that is produced in spontaneous narratives. The study of de Lira et al. shows that, compared to healthy persons, persons with Alzheimer's disease (AD) show more lexical errors (word-finding difficulties, repetitions, and revisions) and a decreased use of coordination to form complex syntactic structures. The model of Orimaye and Golden uses machine learning to train a model using a similar set of variables. Their model, using support vector machines, can distinguish a group with AD and related dementias from a

group with healthy participants with a reported performance (*F*-measure) of 74%. The significant variables reported in their studies are the basis of the variable selection in this article. We apply their models to narratives that are truncated using various cutoff values to quantify the association between text length and uncertainty in the predictive model. The variable sets, language sampling, and predictive tasks of the studies of Orimaye and Golden and de Lira et al. overlap, as do their results. The motivation for using previously described models is that the relationship between text length and predictive accuracy is best measured using variables that are known to be significant for predictions.

Before model training can take place, raw text must be analyzed to obtain measurements for the variables that are used in the model. Software can automate the computing of variables, through the use of automatic parsing and part of speech tagging (Eckhoff & Berdičevskis, 2016; Kübler et al., 2006). The two prior studies by Orimaye and Golden (2014) and de Lira et al. (2011) have successfully used CLAN (MacWhinney, 2009). However, this tool lacks the flexibility to programmatically input text, which is a desired property when experimenting with texts that are truncated at stepwise increasing truncation points. The narratives of one of the experiments reported in this study were analyzed with the Stanford parser rather than with CLAN. In a comparison of parsers for English child–adult dialog language (Huang, 2016), the Stanford parser’s tagging accuracy (95.64%) was comparable to that of CLAN’s MOR tagger (95.59%). A comparative study of probabilistic treebank parsing of German (Kübler et al., 2006) shows that the Stanford parser’s performance for German can be similar to its performance for English, dependent on the training corpus.

We report three different experiments (see Table 1). In the first experiment, the analysis that was used in the prior studies has been repeated to establish baseline results for English, with CLAN as processing tool, as a standard for comparison for the later experiments. The research question is “What is the predictive value of a model that combines the significant linguistic variables from the two prior studies over a model that only includes age as a predictor?” In the second experiment, a corpus of German instead of English was analyzed using the same set of software tools for morphosyntactic analysis and parsing to verify that the same model that works for English can also be applied to German. This establishes a standard of comparison for the third experiment, in which the use of a different processing pipeline, the Stanford parser (Rafferty & Manning, 2008), is evaluated on the German data. The different parser allows analysis of fragments of texts with varying lengths. The comparison between Experiments 1 and 2 establishes the degree to which the model is invariant between English and German, and the comparison between Experiments 2 and 3 establishes the degree to which the calculations for the variables of the model are invariant between CLAN and the Stanford parser.

Bayesian modeling is used to estimate linear regression models. The Bayesian interpretation of probability provides a natural estimation of the relation between model uncertainty

and text length, which is the focus of Experiment 3. A Bayesian model starts with a prior belief of the uncertainty around each parameter of the model, usually uninformative at the start. The probability that a parameter has a given value is updated after each observation, by application of Bayes’ rule. The resulting model yields a probability distribution over the possible mean values of each parameter. The probability distribution quantifies the uncertainty around the estimation of the parameter. An interval can be defined (credible interval) so that it encompasses a given percentage of values. A 95% credible interval represents that, after learning from the data, 95% of the possible values of the parameters of the model will fall within the interval or a 95% probability can be assigned that the parameter estimate is inside the interval. An estimate with more uncertainty will have a wider 95% credible interval than an estimate with less uncertainty.

A credible interval that includes zero indicates that a parameter may contribute nothing to the final prediction. That may warrant further inspection, for example, by stepwise eliminating regressors, if the predictive model itself is the focus of the study. Because we are interested in the uncertainty (credible interval) rather than in the predictive contribution of single parameters, we leave such regressors in the model. If the 95% credible interval does not include zero, then the coefficient is significantly different from 0 and the predictor is important.¹

Experiment 1: Application of the Combined Model of Orimaye and Golden (2014) and de Lira et al. (2011) to English Data with CLAN

Method

Data Set

For this experiment, data from the Pitt corpus of the DementiaBank clinical data set (Becker et al., 1994) were used (see Table 2). The data set has participants with varying diagnoses, split between eight separate groups. In analyses and training, we used data from NBD and (probable/possible) AD groups only, to mirror the groups in the studies of Orimaye and Golden (2014) and de Lira et al. (2011). Elicitations were conducted in English and are based on the description of the Cookie Theft picture component of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983).

¹The Bayesian credibility interval should not be confused with a “confidence interval.” A confidence interval, in its usual frequentist interpretation, is an interval such that the true value of a parameter, when model estimation is repeated, will fall in that interval a given percentage of times. The Bayesian method propagates any lack of information that results from truncation to the posterior probability distribution, leading to a natural estimation of that uncertainty and its relation to fragment lengths.

Table 1. Schematic overview of the three experiments and overall study.

Experiment	Language	Linguistic analysis tools	Goal
Experiment 1	English, Pitt corpus	CLAN	Establish significance of variables reported by Orimaye and Golden (2014) and de Lira et al. (2011)
Experiment 2	German, Jalvingh corpus	CLAN	Establish that analysis is invariant under language change
Experiment 3	German, Jalvingh corpus	Stanford toolchain	Obtain measurements for this study
Overall study			Determine optimal length of texts given a model with predictors

The AD group has a mean age of 71.3 years; the control group (CG) has a mean age of 64.2 years. The difference is statistically significant, as indicated by a Wilcoxon rank sum test ($W = 36898.5$, $p < .01$).

Linguistic Analysis

Data from DementiaBank have been published in a specific form that allows for automatic analysis by a companion program, CLAN (MacWhinney, 2009). In this experiment, a morphosyntactic analysis was first performed using the English grammar of CLAN, and then the EVAL program was used to extract measurements for the relevant variables.

Variables

The variables that were found to be significant by either Orimaye and Golden (2014) or de Lira et al. (2011), or both, were used to compute the regression model:

- age: the age of the participant in years;
- number of utterances: the total number of utterances per narrative;
- mean length of utterances (MLU) in words: the ratio of the total number of words to the number of utterances (Marini et al., 2008);
- number of predicates per utterance: the number of predicates in the discourse per utterance, operationalized as the number of transitive verbs, excluding modals, which are followed by one or more arguments (Surdeanu et al., 2003);
- percentage of verbs: the number of words tagged as verb, participle, copula, or modal as a percentage of the total number of words.
- idea density: measure of propositional idea density; this measure is similar to the one computed by the third major version of the Computerized Propositional Idea Density Rater (Brown et al., 2008), which approximates idea density as the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words;
- number of repetitions: the number of immediate word repetitions in the narrative, as indicated in the transcripts;
- number of revisions: the number of pause positions where a preceding error was retraced to make a correction (Brown et al., 2008).

Independent variables were entered as a block. The values of parameters of the linear regression model and their uncertainty were estimated through Hamiltonian Monte Carlo simulation using STAN (Carpenter et al., 2017) and R.² We computed the model parameters using four chains, each for 20,000 iterations, with the first 10,000 used for warm-up, yielding a total of 40,000 iterations, which were all kept (no thinning). Convergence of the chains was checked using the Brooks–Gelman–Rubin diagnostic (Rhat) and through visual inspection of the diagnostic plots (Gelman et al., 2013).

Results

A diagnostic of the accuracy of the computation of the predictive model is the Rhat value, which represents the consistency between multiple Markov chains. A second diagnostic, the effective sample size per iteration (ESS; N_{eff} / N), is used to quantify whether the number of iterations is large enough for inferencing. Common guidelines dictate that Rhat values approximate 1 and that the ESS values exceed 0.001.

In the model trained for this experiment, the Rhat values for each of the parameters were approximately 1.00, indicating that all the chains in the Monte Carlo simulation converged. The minimum ESS always exceeded the threshold of 0.001, indicating that the chains were of sufficient length. The RStan function to check diagnostics (`check_hmc_diagnostics`) reported that the n_{eff} / N measure looks reasonable for all parameters, that zero of 40,000 iterations ended with a divergence (0%), that zero of 40,000 iterations saturated the maximum tree depth of 10 (0%), and that the E-BFMI (Estimated Bayesian Fraction of Missing Information) indicated no pathological behavior.

The resulting estimated model parameters are reported in Table 3. The parameters age, number of utterances, MLU in words, number of repetitions, and number of revisions each have credible intervals that do not contain zero and hence contribute to the model.

The comparison of model performance in Table 4 shows the average Akaike weight for each model, which is an estimate of the probability that the model will make the best predictions on new data, conditional on the set of models considered (McElreath, 2016, p. 199). The top model in Table 4 receives a weight of 1, indicating that, based on the

²R v3.3.3 (R Core Team, 2017), with packages rethinking (McElreath, 2016) and mc-stan (Stan Development Team, 2018).

Table 2. Demographics of participants from the Pitt corpus (Becker et al., 1994) and from the Jalvingh corpus.

Variable	Pitt corpus		Jalvingh corpus	
	Control group (<i>n</i> = 101) <i>M</i> (<i>SD</i>) Range	AD group (<i>n</i> = 181) <i>M</i> (<i>SD</i>) Range	Control group (<i>n</i> = 8) <i>M</i> (<i>SD</i>) Range	AD group (<i>n</i> = 9) <i>M</i> (<i>SD</i>) Range
Age (years;months) ^a	63;8 (8;3) 46.2–81.9	71;4 (8;3) 50.0–88.7	75;1 (8;0) 64;4–88;5	78;8 (8;6) 70;3–87;1
Education (years)	14.3 (2.9) 8–20	12.1 (2.9) 8–20	11.7 (1.9) 9–14	11.8 (1.6) 10–14
Mini-Mental State Examination	29.1 (1.1) 26–30	18.4 (5.2) 8–30	29.9 (0.35) 29–30	20.8 (2.6) 16–24
Sex (male/female)	44/57	59/121	2/6	5/4
Handedness (R/L/Ambi)	93/7/1	172/4/5	8/0/0	7/0/2

Note. For the Jalvingh corpus, control participants were interviewed 3 times in one session, and participants with Alzheimer's disease (AD) were interviewed 3 times each in three different sessions. R/L/Ambi = right/left/ambidextrous.

^aAD participants in the Jalvingh corpus were seen multiple times. The age at first participation is used for reporting.

computed information criterion, the model, which includes linguistic variables, has the best out-of-sample accuracy.

The area under receiver operating characteristic (ROC) curve plot (cf. Figure 1) of the performance of both models visualizes the trade-off between true and false positive rates—at any rate of false positives, the true positives predicted by the model with linguistic predictors outnumber the true positives predicted by the age-only model. The total area under curve in an ROC plot represents the probability that the model predicts a higher value for a given random positive example than for a random negative example. There is a clear difference between the two models. The area under the ROC curve of the model with only age as predictor is 71.84%, and that of the model with both age and linguistic predictors is 81.41%. The classifier accuracy of the model with age and linguistic variables, measured as F1 score, is 78.6%, with precision 75.1% and recall 82.4%. This accuracy is close to the reported F1 score (74%) in the study of Orimaye and Golden (2014). The confusion matrix for the classifier is reported in Table 5.

Table 3. Estimated model parameters, trained on the Pitt corpus, analyzed with CLAN.

Variable	<i>M</i>	<i>SD</i>	Lower 0.95	Upper 0.95
Intercept	0.76	5.92	−10.64	11.90
Age	−0.11	0.01	−0.14	−0.09
No. of utterances	0.18	0.03	0.11	0.24
MLU in words	0.26	0.08	0.11	0.41
No. of predicates per utterance	−0.40	0.43	−1.27	0.39
Percentage of verbs	0.00	0.04	−0.07	0.08
Idea density	−0.45	0.78	−1.90	1.05
No. of repetitions	−0.43	0.11	−0.64	−0.22
No. of revisions	−0.23	0.08	−0.38	−0.09

Note. Lower/higher 0.95 indicates the credible intervals, sampled from the posterior distribution. MLU = mean length of utterances.

Experiment 2: Application of the Model to German Data Processed With CLAN

Method

Data Set

Spoken language was elicited from German speakers with NBD (*n* = 8) and individuals with a clinical diagnosis of AD (*n* = 9). The data form a subset of a corpus collected for a larger study of processing of verbs and nouns in speakers with different types of dementia (partially published as the *Jalvingh DementiaBank Corpus* [Jalvingh, 2016]; henceforth, Jalvingh corpus). Participant recruitment and data elicitation were performed in the context of that study. In this experiment, only the transcripts were included from participants assigned to the CG and the AD group of the Jalvingh corpus, so that the study population aligns with the populations in the studies of Orimaye and Golden (2014) and de Lira et al. (2011).

Persons in the CG participated once, resulting in three different narratives. To follow the progression of their disease, persons in the AD group participated at three separate moments, each session about 6 months apart, resulting in a maximum number of nine different narratives, one for each topic, repeated 3 times.

Participants were interviewed on the topics “past,” “present,” or “future.” Participants were asked to speak of childhood memories (topic: past), of a typical day in daily life (topic: present), and of plans that they may have for the next week, month, or year (topic: future).

The interviewer did not follow a set script with prompts but rather introduced the topic and invited the participant to speak. The interviewer used intermediate feedback (“Aha,” “Oh really?”) and sometimes asked targeted questions based on the narrative to keep the conversation moving. Transcriptions of the interviewer's language were excluded a priori. The interview ended after 10 min or when the participant was finished.

Table 4. Comparison statistics of models trained on Pitt corpus, analyzed with CLAN.

Variable	WAIC	pWAIC	dWAIC	Weight	SE	dSE
Age + linguistic variables as predictors	491	8.6	0	1	21.7	NA
Age as only predictor	561	1.4	70	0	12.8	17
Intercept-only model	623	1.0	132	0	8.7	21

Note. WAIC = Watanabe–Akaike information criterion; pWAIC = effective number of parameters; dWAIC = delta with model with lowest out-of-sample deviance; Weight = estimated weight if used in ensemble (Akaike); SE/dSE = standard error, delta with highest ranking model.

Patients were diagnosed with neurodegenerative diseases according to German neurological standards (Leitlinien Deutsche Gesellschaft für Neurologie). The diagnosis was the criterion for classification into AD group or CG. The participants were matched according to age and education level. Exclusion criteria were depression as assessed with the short version of the geriatric depression scale (Yesavage et al., 1982) or severe visual and hearing problems. The Mini-Mental State Examination (MMSE; Folstein et al., 1975) was used for the assessment of the severity of the disease. Ethical approval was assigned by the medical ethics committee Hannover.

The AD group has a mean age of 75 years; the CG has a mean age of 73.4 years. The difference is not statistically significant as indicated by a Wilcoxon rank sum test ($W = 839$, $p = .64$).

The CG ($n = 8$ participants) has a mean MMSE score of 29.88. Due to experimental attrition, the number of participants decreased between the interview moments. The AD group has mean MMSE scores at interview Moment 1 of 22.11 ($n = 9$ participants), Moment 2: 20.25 ($n = 8$ participants), and Moment 3: 19.25 ($n = 4$ participants). The Wilcoxon rank sum test indicated that the MMSE score was statistically significantly lower for the AD group than for the CG ($W = 72$, $p < .01$) at alpha = .05 at Moment 1,

but not within that group between interview Moments 1 and 2 ($W = 54$, $p = .09$) or Moments 2 and 3 ($W = 18$, $p = .8$).

Linguistic Analysis

As in Experiment 1, the CLAN tools were used for morphosyntactic analysis and variable measurements. In this experiment, the German grammar of CLAN was used.

Variables

The same variables as in Experiment 1 were used: age, number of utterances, MLU in words, number of predicates per utterance, percentage of verbs, propositional idea density, number of repetitions, and number of revisions. As in Experiment 1, parameters were entered as a block in the regression, and their values with their associated uncertainty were then estimated through Hamiltonian Monte Carlo simulation using STAN and R. We computed the model parameters using four chains, each for 20,000 iterations, with the first 10,000 used for warm-up, yielding a total of 40,000 iterations, which were all kept (no thinning). Convergence was checked using the Brooks–Gelman–Rubin diagnostic (Rhat) and through visual inspection of the diagnostic plots (Gelman et al., 2013).

Results

The sample lengths for the German JALVINGH data are higher than those of the English Pitt data. In addition, the words-per-minute measure in the German data set is much higher than the one in the Pitt data set (for controls: 120.06 in the Pitt data set vs. 203.16 in the German data set; cf. Figure 2).

The one-way between-groups analysis of variance showed that there is no meaningful group difference between

Figure 1. Model performance of models with and without linguistic predictors applied to English data. Area under receiver operating characteristic curve for the age-only and age + language models applied to the Pitt data. **Table 5.** Confusion matrix of the classifier (dependent variable: “membership of control group”) trained on the Pitt corpus in Experiment 1.

Prediction	Reference	
	False	True
False	230	75
True	48	109

Figure 2. The sample length and words per minute of English and German samples. Box plots of the variables “sample length” and “words per minute” in the English and German samples. En = English; Ge = German; AD = Alzheimer’s disease; p/m = per minute; wpm = words per minute.

the participants in the German data for the variable sample length at the level of $p < .05$, $F(1,86) = 0.81$, $p = .37$. The one-way between-groups analysis of variance showed that there is no meaningful group difference between the participants in the German data for the variable words-per-minute measure at the $p < .05$ level, $F(1,86) = 1.82$, $p = .18$.

In the model trained for this experiment, the Rhat values for each of the parameters were approximately 1.00, indicating that all the chains in the Monte Carlo simulation converged. The minimum ESS always exceeded the threshold of 0.001, indicating that the chains were of sufficient length. The RStan function to check diagnostics (`check_hmc_diagnostics`) reported that the n_{eff} / N measure looks reasonable for all parameters, that zero of 40,000 iterations ended with a divergence (0%), that zero of 40,000 iterations saturated the maximum tree depth of

10 (0%), and that the E-BFMI indicated no pathological behavior.

The estimated model parameters are reported in Table 6. The parameter “percentage of verbs” has a credible interval that does not contain zero and hence contributes to the model. Compared on the out-of-sample performance using Akaike information criteria (cf. Table 7), the model with linguistic variables has a weight of 1 versus the age-only model with a weight of 0. This implies that, based on these criteria, the probability that the model with linguistic variables predicts better than the age-only model is 1.

Because the data set is highly skewed, with 62 narratives by the AD group and 24 by the CG, the F1 score is uninformative to report classifier accuracy. The ROC curve gives an estimation of the classifier’s performance under various label decision thresholds and is more useful for imbalanced data sets. As with the application of this model to

Table 6. Estimated model parameters of the fitted model, trained on the Jalvingh corpus, analyzed with CLAN.

Variable	<i>M</i>	<i>SD</i>	Lower –0.95	Upper –0.95
Intercept	0.83	7.69	–13.20	17.29
Age	0.04	0.04	–0.04	0.13
No. of utterances	0.01	0.01	–0.01	0.02
MLU in words	–0.21	0.12	–0.45	0.03
No. of predicates per utterance	0.26	0.94	–1.43	2.20
Percentage of verbs	–0.93	0.25	–1.38	–0.41
Idea density	–0.04	0.72	–1.32	1.42
No. of repetitions	–0.04	1.01	–2.20	1.73
No. of revisions	0.00	1.04	–2.02	2.05

Note. MLU = mean length of utterances.

the Pitt data, the area under the curve (cf. Figure 3) for the model that includes linguistic variables is visibly larger than the age-only model. The area under the ROC curve of the model with only age as predictor is 62%.

Regardless of the false positive rate, it predicts only little over 50%. This reflects that, in the Jalvingh corpus, participants in both groups have similar ages. The area under ROC curve of the model with both age and linguistic predictors is 78%.

Experiment 3: Application of the Model to German Data Processed With Stanford Toolchain Methods With Text Length Variations

Method

Data Set

In this experiment, the same data set was used as in Experiment 2, conversations in German with German speakers with NBD ($n = 8$) and individuals with a clinical diagnosis of a form of dementia: (probable) AD ($n = 9$). Different training sets were construed by truncating the narratives at a set text length. Texts were truncated using the following strategy: Given a cutoff value of n words, a text is truncated at the last terminal node that is still part of the parse tree (S-node) that contains terminal node n . This ensures that the truncated text contains at least n words and that truncation does not cause interrupted sentences. The

truncation values that were used were 100, 300, 500, ... 1,500 words.

Linguistic Analysis

A toolchain was used that allows easy experimentation with varying text lengths. We used the German grammar of the Stanford parser (Rafferty & Manning, 2008) with postprocessing of the results by custom scripts in R.

Variables

The linguistic variables (in parentheses: the abbreviated names of the variables) used in the model for Experiment 3 are as follows:

- Lexical level
 - Percentage of verbs (pct.verbs): percentage of verbs (verb, participle, copula, modal) on all words.
 - Density (density): variable of propositional idea density (Brown et al., 2008), which is the sum of (adjectives, prepositions, adverbs, cardinals, nouns, and verbs) divided by the total number of words.
 - Number of repetitions (# repetitions): a variable for the number of repetitions of adjacent lexical items, operationalized as the number of runs of equal terminal nodes divided by the total number of terminal nodes.
- Syntactic level
 - MLU (MLU.Utts): mean length of utterances in the fragment.

Table 7. Comparison statistics of models trained on Jalvingh corpus, analyzed with CLAN.

Variable	WAIC	pWAIC	dWAIC	Weight	SE	dSE
Age + linguistic variables as predictors	94	4.5	0	1	11.0	NA
Age as only predictor	107	1.1	13	0	8.2	8.7

Note. WAIC = Watanabe–Akaike information criterion; pWAIC = effective number of parameters; dWAIC = delta with model with lowest out-of-sample deviance; Weight = estimated weight if used in ensemble (Akaike); SE/dSE = standard error, delta with highest ranking model.

Figure 3. Model performance of models with and without linguistic predictors applied to German data. Area under receiver operating characteristic curve for the age-only and age + language models applied to the German data from the Jalvingh corpus.

- MLU in words (MLU.Words): average number of lexical items per utterance.
- Syntactic index (si.index): a measure for the syntactic complexity of the discourse; the ratio of complex sentences (coordinated sentences, subordinate and relative clauses) to the total number of sentences (Duong & Ska, 2001).
- Verbs per utterance (Verbs.Utt): average number of finite verb forms per sentence, excluding modals. Approximation of the number of clauses per utterance.

The linear regression model used in this experiment is similar to the model in Experiments 1 and 2, but the variables density and number of repetitions were computed in a different way: The density variable was computed as originally proposed by Brown et al. (2008), and the number of repetitions was computed as the proportion of immediately adjacent unique lexical items compared to the total number of lexical items.

Statistical Analysis

Because CLAN was used in the first two experiments, those variables from the sets of Orimaye and Golden (2014) and de Lira et al. (2011) were excluded that cannot be computed directly using CLAN, that is, number of coordinates and number of reduced sentences. The resulting regression model is compared to a model that includes only age as predictor to quantify how a model with linguistic variables

fares better predicting group membership than a model without linguistic variables.

The model maps parameters onto a binomial distribution using the logit link function. All parameters (except age) were normalized to the mean and scaled around their standard deviation. To reflect a lack of prior information, uninformative priors were specified for the parameters of the model using a normal distribution with a mean of zero and a large standard deviation of 1,000.

The parameters of the logistical regression model that was used are presented as odds ratio (O/R, logodds). One model is presented as the base model in Figure 4, as an example of the shape of the models we used.

As in Experiments 1 and 2, parameters were entered as a block in the regression, and their values with their associated uncertainty were then estimated through Hamiltonian Monte Carlo simulation using STAN and R. We computed the model parameters using four chains, each for 20,000 iterations, with the first 10,000 used for warm-up, yielding a total of 40,000 iterations, which were all kept (no thinning). Convergence was checked using the Brooks–Gelman–Rubin diagnostic (Rhat) and through visual inspection of the diagnostic plots (Gelman et al., 2013).

The predictive accuracy of the Bayesian models is evaluated using information criteria (Akaike, 1974). One measure that allows comparison between models is the Watanabe–Akaike information criterion (or widely applicable information criterion; Watanabe, 2010), which is an approximation of the predictive accuracy, as measured by out-of-sample deviance.

The resulting Bayesian model yields a posterior. A posterior is a set of estimates of the relative plausability of different parameter values, conditional on the data that was used for training. The posterior is used to generate a posterior distribution, a set of simulated data points that can be inspected for the uncertainty in the estimation of parameter values. We use the posterior distribution to quantify the uncertainty associated with various text lengths.

We first establish the likelihood that a participant has AD or NBD (p_0) using a model that takes only age into account. In the model with age and linguistic variables, we consider a parameter significant if and only if it influences the final prediction so much that the predicted likelihood p_1 differs more from p_0 than the margins of uncertainty around p_1 and p_0 . This decision procedure is equivalent to the decision procedure approach of Kruschke (2014) using a region of practical equivalence (ROPE). If the margin of uncertainty associated with a model parameter grows, then measurements of that variable must be further apart to become statistically significant. Assuming that there is a relationship between text length and information entropy, hence parameter uncertainty, the significance of model parameters under varying text lengths can be used to estimate text length requirements for various variables.

Results

In the models trained for this experiment, the Rhat values for each of the parameters were approximately 1.00,

Figure 4. The base model used to predict the uncertainty around regressors.

isControl	~	Binomial($n=1, p$)	(likelihood)
logit(p)	=	a $+ \beta_{\text{age}} * \text{age}$ $+ \beta_{\text{MLU.Utts}} * \text{MLU utterances}$ $+ \beta_{\text{MLU.Words}} * \text{MLU Words}$ $+ \beta_{\text{density}} * \text{density}$ $+ \beta_{\text{predicatesPerUtterance}} * \# \text{ predicates per utterance}$ $+ \beta_{\text{pct.verbs}} * \text{percentage verbs}$ $+ \beta_{\text{repetitions}} * \# \text{ repetitions}$ $+ \beta_{\text{si.index}} * \text{SI-index}$	(linear model)
a	~	Normal($\mu = 0, \sigma = 1000$)	(prior for intercept)
$\beta_{\text{age}}, \beta_{\text{MLU.Utts}}$ (etc.)	~	Normal($\mu = 0, \sigma = 1000$)	(priors for regressors)

indicating that all the chains in the Monte Carlo simulation converged. The minimum ESS always exceeded the threshold of 0.001, indicating that the chains were of sufficient length. The RStan function to check diagnostics (`check_hmc_diagnostics`) reported that the n_{eff} / N measure looks reasonable for all parameters, that zero of 40,000 iterations ended with a divergence (0%), that zero of 40,000 iterations saturated the maximum tree depth of 10 (0%), and that the E-BFMI indicated no pathological behavior.

The toolchain of Experiments 1 and 2 (CLAN) measures variables in a different way than the toolchain for Experiment 3 (Stanford parser). The advantage of the Stanford parser toolchain over the use of CLAN is that it is much easier to programmatically feed texts. CLAN lacks a programmatic interface. The programmatic interface is needed to compute variables while truncating the texts at various truncation points.

Comparing the two toolchains, there is a significant correlation between the computed measures for MLU in words ($r = 1, p < .001$), MLU in utterances ($r = .99, p < .001$), and percentage of verbs ($r = .48, p < .001$), but not for the measure of density ($r = -.17, p \geq .05$). The reason why the percentage of verbs is different between the CLAN and the Stanford parser is the result of differences in the calculation of sentence length between the parsers: the Stanford parser counts all terminals (words and punctuation marks [.] and [?]), whereas CLAN counts only words.

The trained model had factors with a high variance inflation factor (VIF; > 10), as computed by the VIF function of the car package (Fox & Weisberg, 2011). This suggests that two or more variables are multicollinear. After stepwise regression, we eliminated the verbs per utterance variable, resulting in a model with all VIF scores for all variables less than 3.1.

The revised model, trained on variables computed by the alternative toolchain, performs better than the model trained on variables computed by CLAN (cf. Figure 5). The area under the curve of the CLAN variables model is

78%, and that of the revised model with variables computed by the Stanford parser is 90%.

Analysis of Relation Between Text Length and Predictor Parameter Uncertainty

The estimated model parameters for the full narratives, without truncation, are reported in Table 8. Significant variables are number of utterances, MLU in words, idea density, number of repetitions, and syntactic index.

Figure 5. Model performance of models trained with CLAN annotations and trained with Stanford toolchain annotations. Area under receiver operating characteristic curve for models trained on German narratives that were trained with CLAN or with the Stanford toolchain.

Table 8. Estimated model parameters of the fitted model, trained on full (nontruncated) fragments of the Jalvingh corpus, and analyzed with the Stanford parser.

Variable	<i>M</i>	<i>SD</i>	Lower 0.95	Upper 0.95
Intercept	-7.43	4.25	-14.28	-0.83
Age	0.07	0.05	-0.02	0.15
No. of utterances	0.80	0.46	0.06	1.54
MLU in words	5.26	2.23	1.82	8.93
No. of predicates per utterance	0.45	0.48	-0.32	1.22
Idea density	2.64	0.73	1.28	3.82
No. of repetitions	-3.52	0.97	-5.04	-1.97
Syntactic index	-5.40	2.33	-9.07	-1.68

Note. MLU = mean length of utterances.

Different models were trained, each under different truncation conditions. To illustrate the interpretation of the results, consider the results for variable density, as plotted in Figure 6. The age-only model predicts that, at the age of 74 years (the average age in the corpus), the probability of any participant to be a speaker with NBD is between .18 and .37. This is plotted in Figure 6 as the red shaded area, with the top and bottom borders at these probabilities. The probabilities that a participant is a speaker with NBD at varying values of the density variable are plotted by the blue line, with the uncertainty at each point represented by the gray-shaded area. If trained on whole fragments (upper left panel of Figure 6), the model predicts that when the density is 1 *SD* more than the mean, with all other variables at their means, the likelihood of the participant being a speaker with NBD is estimated in the posterior to be between $p = .54$ and $p = 1$, with as mean likelihood $p = .54$. In order for the density variable to be so influential that the final prediction of the model is significantly different, the final model prediction margins of uncertainty should clear the uncertainty margins of the age-only prediction model. In terms of visualization, this is where the gray-shaded area ceases to overlap the red-shaded area in Figure 6. In the case of the variable density, this occurs at $x = 2.58$ *SDs*, which corresponds to a true density value of 0.66. The other panels of Figure 6 show the same variable, but now trained on texts of 1,100, 500, and 300 words in length. The distance between the blue prediction line and the end of the uncertainty bound, at the top of the age-only band, is a measure of the “required deviation of a single variable for a model to become informative” (ROPE distance). The measure varies when the model is trained with longer or shorter text fragments; this is illustrated in the upper left plot as Δ .

The uncertainty around the variable decreases when machine learning is performed on longer fragments. The relation between text length and variable significance margins is plotted in Figure 7. The true variation in the German data is 3.94 *SDs* from the mean, indicated by the blue dashed lines. In general, the uncertainty decreases sharply, up to at about 700 words, where the uncertainty

starts to stabilize. There are two exceptions. The variable MLU (in words, scaled) shows less decline in required standard deviation, and the decline starts later, at texts of length of 1,100. The variable ratio of verbs (%) shows no decline at all.

Discussion

The model derived from the studies of Orimaye and Golden (2014) and de Lira et al. (2011) was applied to English in the first experiment and to German in the second experiment. It differed which predictors were found to be significant. When trained on English data in Experiment 1, the predictors with a credible interval beyond zero are age, number of utterances, MLU in words, number of repetitions, and number of revisions. When trained on German data in Experiment 2, the single predictor with a credible interval beyond zero is percentage of verbs. A possible explanation for the difference of these outcomes is that the data set used for the model in Experiment 1 contained more data points, with shorter narratives elicited in a different way than the data set for the model of Experiment 2. This leads to more uncertainty around the predictors that are based on text length. We speculate that participants are less error-prone (predictors number of repetitions and number of revisions) if the task elicits speech that is more spontaneous. It can be argued that the elicitation task in Experiment 2 is more spontaneous, because participants in that experiment had more freedom to choose the content of the conversation.

The model trained in Experiment 3 is the best predicting model. Because the only variation between Experiment 2 and Experiment 3 is the different analysis toolchain, we hypothesize that the way in which we compute the variables leads to better predictive results. In particular, the measures of length (number of utterances/MLU in words) are significant in the first and third experiments and in prior studies. We believe that the way CLAN computes these variables for German texts warrants more investigation.

The results of the second experiment suggest that, taken as a block, a machine learning model trained on the variables proposed by Orimaye and Golden (2014) and de Lira et al. (2011) is equally predictive for German as for English language fragments. This suggests that the uncertainty around the predictors is related to properties of the language of the participants, one of which is text length. The use of the Stanford parser, rather than CLAN, improves the performance of the model. The density variable as computed in this study is more predictive than the density variable that is computed by CLAN. The results of the third experiment suggest that the uncertainty around single variables can be reduced by using longer text fragments to train the model.

In general, uncertainty in machine learning trained model can decrease when (a) differences between measurements become larger, when (b) the certainty of individual measurements increases, or when (c) the amount of training

Figure 6. The influence of the density variable on the predicted probability of Alzheimer's disease (AD). The prediction that a participant is a speaker with no brain damage given a normalized value for the density variable and all other variables constant at default levels. In the upper left plot, the model used for prediction was trained on whole fragments; the other plots show models where texts were truncated at various points. Between the dotted blue lines (ROPE boundary), the prediction is practically equivalent to the prediction based on an age-only model. Beyond this zone (ROPE distance), the prediction differs significantly. CI: Credible Interval.

data available increases. In our context, increasing text length or collecting data from more individuals may be used as strategies for (b) and (c). This article studied (b). There is a similar pattern of decreasing uncertainty

for four of the six variables. In the specific case of this corpus, observations in a range of 0–4 SDs were observed. A model with predictor variables that deviate within that range can be trained with 500 words or more. When trained

Figure 7. The relation between text length and the uncertainty around a regressor. Variation required to influence prediction for various text lengths. Each point represents how far (max: 16 SDs) the variable (*ceteris paribus*) must deviate from the mean to influence the final model prediction. The dashed line represents the standard deviation as observed in the original data, with 5% outliers removed. MLU = mean length of utterances.

on shorter texts, the uncertainty around the predictor variables becomes so large that measurements would have to have unrealistic deviations in order to become significant, at least in the context of this corpus. However, the informative gain starts to diminish when sample length increases beyond 1,100 words. Variables that are not significant in prediction at all, such as ratio of verbs in Figure 7,

show to be insensitive to text length modulations because, regardless of text length, their uncertainty remains equal.

It is surprising how much the different variables converge to become predictive at the same cutoff in terms of sample length, at around 700 words. The blue dashed lines in Figure 7 denote the observed variation in the actual data—each of the variables has an observed variance of about

3.5 SDs. This finding is remarkably consistent among the variables and suggests that, after scaling, each of the variables has a very similar bell curve. Future research may indicate whether there is a relation between the variance of a variable and how much data are needed to predict that variable, given an acceptable degree of uncertainty.

Of course, the essence of linear regression models is that they use multiple regressors to make a prediction. The focus in this study was on the implied predictions of single variables in isolation. Varying one variable while keeping the others constant may be misleading because, in reality, variables never change in isolation. The uncertainty around the joint parameters will be less, because of the additive effect of multiple variables. This also prohibits interpreting any of the results as valid probability values, because all predictors except the one of interest are given default values. We compute uncertainty under the assumption that a participant is 74 years old (the population's mean age) and that all other measurements are at their mean average. Under any other assignment of variables, the location (but not the shape) of the curve in Figure 7 may change.

One weakness of the strategy in this article is that it assumes that the distribution of measurements over a fragment is uniform: that there is no difference between a measurement taken at the start of a text and one taken at any other position in the text. Thus, we do not take into account that some speakers, for example, have difficulties while starting to speak, which do not manifest later in a conversation.

Conclusions and Further Work

In practice, the guideline for determining how much data to collect is “as much as possible.” The current study gives a reasoned and quantified answer yielding both a lower and an upper limit, but only for a specific situation. When other research would target a different population or a different set of variables, the uncertainty would have to be computed again. In this case, a text length of 700 words becomes reliably predictive to tell AD and NBD apart, if a single variable model is used. The addition of more text over 900 words does not add more certainty to the logistic regression model in this case study. It is senseless to “collect as much as possible” beyond this threshold. The results are used to establish the protocol for the researchers who perform elicitations for the larger longitudinal study that forms the context of this research. The computed numbers are used as lower and upper bounds of text length; samples within these bounds are both predictive enough for research purposes and not longer than necessary.

The model used in this study was fixed to the models reported previously in the literature. An extension of this study could involve other variables as well to see if they behave similarly. We hypothesize that longer text samples are needed if more higher level variables are included in a predictive model.

The computed optimal text length varies with the population. When differences between unimpaired and

impaired speakers become more pronounced, smaller text samples will suffice for diagnosis. The method in this article could be extended to provide a quantitative measure of the degree to which populations differ in terms of their language production: If the result of the computation is that a large text length is needed, between-groups variation of language may be small, and vice versa. As the difference between impaired and unimpaired increases, the computation will predict smaller sample sizes. The computed sample size may thus be used as a measure of distance between two groups.

This study was limited to German speakers. It is an open question whether the stable finding of about 700 words as optimal sample length extends to other languages as well.

Although the results of this study are not directly clinically applicable, our method may be interesting to other researchers who collect discourse data for clinical purposes in other populations. There are some suggestions for further research that do have clinical implications. First, the way in which the uncertainty estimation converges to a point is a reflection of the idea that there is an optimal maximum length for narratives produced in a spontaneous speech take. This optimal maximum length may provide a minimum bound for a clinician when evaluating a language sample for diagnostic purposes. Second, the difference in the significance of the parameters between the models for Experiment 1 and that for Experiment 2 suggests that the type of elicitation (spontaneous vs. semispontaneous) has implications for the variables that should be used in a prediction model. Some speech tasks can be completed with narratives that are relatively short. If a given model requires longer text samples, the protocol could include speech tasks that typically elicit longer fragments.

Acknowledgments

This work is part of the research program “Doctoral Grant for Teachers,” with Project 023.007.061. Roelien Bastiaanse is partially supported by the Center for Language and Brain NRU Higher School of Economics, RF Government Grant 14.641.31.0004. We thank Casper Albers and Gertjan van Noord for helpful discussions of early versions of this article.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Becker, J. T., Boiler, 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. <https://doi.org/10.1001/archneur.1994.00540180063015>
- Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, 8, 269. <https://doi.org/10.3389/fpsyg.2017.00269>
- Brookshire, R. H., & Nicholas, L. E. (1995). Relationship of word frequency in printed materials and judgments of word frequency in daily life to Boston Naming Test performance of aphasic adults. *Clinical Aphasiology*, 23, 107–119.

- Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540–545. <https://doi.org/10.3758/BRM.40.2.540>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Çokal, D., Sevilla, G., Jones, W. S., Zimmerer, V., Deamer, F., Douglas, M., Spencer, H., Turkington, D., Ferrier, N., Varley, R., Watson, S., & Hinzen, W. (2018). The language profile of formal thought disorder. *NPJ Schizophrenia*, 4(1), 18. <https://doi.org/10.1038/s41537-018-0061-9>
- de Lira, J. O., Ortiz, K. Z., Campanha, A. C., Bertolucci, P. H. F., & Minett, T. S. C. (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23(3), 404–412. <https://doi.org/10.1017/S1041610210001092>
- Duong, A., & Ska, B. (2001). Production of narratives: Picture sequence facilitates organizational but not conceptual processing in less educated subjects. *Brain and Cognition*, 46(1–2), 121–124. [https://doi.org/10.1016/S0278-2626\(01\)80047-6](https://doi.org/10.1016/S0278-2626(01)80047-6)
- Eckhoff, H. M., & Berdičevskis, A. (2016). Automatic parsing as an efficient pre-annotation tool for historical texts. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities* (pp. 62–70). The COLING 2016 Organizing Committee.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Minimal state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189–198. [https://doi.org/10.1016/0022-3956\(75\)90026-6](https://doi.org/10.1016/0022-3956(75)90026-6)
- Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Sage <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.
- Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55, 43–60. <https://doi.org/10.1016/j.cortex.2012.12.006>
- Fraser, K. C., Rudzicz, F., & Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. *Interspeech*, 2177–2181.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman & Hall/CRC.
- Goodglass, H., & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination booklet*. Lea & Febiger.
- Grande, M., Hussmann, K., Bay, E., Christoph, S., Piefke, M., Willmes, K., & Huber, W. (2008). Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia. *International Journal of Language & Communication Disorders*, 43(4), 408–426. <https://doi.org/10.1080/13682820701685991>
- Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, 3(5), 1225–1241. <https://doi.org/10.1111/j.1749-818X.2009.00149.x>
- Hinzen, W., Rosselló, J., Morey, C., Camara, E., Garcia-Gorro, C., Salvador, R., & de Diego-Balaguer, R. (2018). A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage Huntington's disease. *Cortex*, 100, 71–83. <https://doi.org/10.1016/j.cortex.2017.07.022>
- Huang, R. (2016). *An evaluation of POS taggers for the CHILDES corpus* [Master's thesis, The Graduate Center, City University of New York]. CUNY Academic Works. https://academicworks.cuny.edu/gc_etds/1577/
- Hussmann, K., Grande, M., Meffert, E., Christoph, S., Piefke, M., Willmes, K., & Huber, W. (2012). Computer-assisted analysis of spontaneous speech: Quantification of basic parameters in aphasic and unimpaired language. *Clinical Linguistics & Phonetics*, 26(8), 661–680. <https://doi.org/10.3109/02699206.2012.683672>
- Jalvingh, F. (2016). *DementiaBank Jalvingh Corpus* [Data set]. TalkBank. <https://doi.org/10.21415/t5kw4t>
- Kübler, S., Hinrichs, E. W., & Maier, W. (2006). Is it really that difficult to parse German? *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 111–119). <https://doi.org/10.3115/1610075.1610093>
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
- MacWhinney, B. (2009). *The CHILDES Project Part 2: The CLAN programs*. Department of Psychology.
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Springer. <https://doi.org/10.1057/9780230511804>
- Marini, A., Spoletini, I., Rubino, I. A., Ciuffa, M., Brià, P., Martinotti, G., Banfi, G., Buccascino, R., Strom, P., Siracusano, A., Caltagirone, C., & Spalletta, G. (2008). The language of schizophrenia: An analysis of micro and macrolinguistic abilities and their neuropsychological correlates. *Schizophrenia Research*, 105(1–3), 144–155. <https://doi.org/10.1016/j.schres.2008.07.011>
- McElreath, R. (2016). *Rethinking: Statistical rethinking book package*.
- Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive Science*, 42(Suppl. 2), 375–412. <https://doi.org/10.1111/cogs.12592>
- Nakai, Y., Takiguchi, T., Matsui, G., Yamaoka, N., & Takada, S. (2017). Detecting abnormal word utterances in children with autism spectrum disorders: Machine-learning-based voice analysis versus speech therapists. *Perceptual and Motor Skills*, 124(5), 961–973. <https://doi.org/10.1177/0031512517716855>
- Orimaye, S. O., & Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In P. Resnik, R. Resnik, & M. Mitchell (Eds.), *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 78–87). Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-3210>
- Pakhomov, S. V., Smith, G. E., Marino, S., Birnbaum, A., Graff-Radford, N., Caselli, R., Boeve, B., & Knopman, D. S. (2010). A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23(2), 127–144. <https://doi.org/10.1016/j.jneuroling.2009.12.001>
- Peintner, B., Jarrold, W., Vergyriy, D., Richey, C., Tempini, M. L. G., & Ogar, J. (2008). Learning diagnostic models using speech and language measures. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008* (pp. 4648–4651). <https://doi.org/10.1109/IEMBS.2008.4650249>
- Prins, R., & Bastiaanse, R. (2004). Analyzing the spontaneous speech of aphasic speakers. *Aphasiology*, 18(12), 1075–1091. <https://doi.org/10.1080/02687030444000534>
- R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software]. <https://www.R-project.org/>
- Rafferty, A. N., & Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German* (pp. 40–46). <https://doi.org/10.3115/1621401.1621407>

- Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J.** (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), 2081–2090. <https://doi.org/10.1109/TASL.2011.2112351>
- Stan Development Team.** (2018). *Stan Modeling Language Users Guide and Reference Manual (Version 2.18.0)* [Computer software]. <http://mc-stan.org>
- Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P.** (2003). Using predicate-argument structures for information extraction. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, 8–15. <https://doi.org/10.3115/1075096.1075098>
- Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M.** (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7, 195. <https://doi.org/10.3389/fnagi.2015.00195>
- Vermeulen, J., Bastiaanse, R., & Van Wageningen, B.** (1989). Spontaneous speech in aphasia: A correlational study. *Brain and Language*, 36(2), 252–274. [https://doi.org/10.1016/0093-934X\(89\)90064-3](https://doi.org/10.1016/0093-934X(89)90064-3)
- Watanabe, S.** (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116), 3571–3594.
- Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M.** (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE*, 13(2), e0192773. <https://doi.org/10.1371/journal.pone.0192773>
- Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O.** (1982). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research*, 17(1), 37–49. [https://doi.org/10.1016/0022-3956\(82\)90033-4](https://doi.org/10.1016/0022-3956(82)90033-4)
- Yngve, V.** (1961). The depth hypothesis. In R. Jakobson (Ed.), *Structure of Language and its Mathematical Aspects* (pp. 10–136). American Mathematical Society. <https://doi.org/10.1090/psapm/012/9974>
- 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(3), 1145–1160. <https://doi.org/10.3233/JAD-160099>