university of groningen / faculty of arts

Netherlands Organisation for Scientific Research

UNIVERSITY OF APPLIED SCIENCES UTRECHT

Determining the ideal length of spontaneous speech fragments for predictive analysis

Roelant Ossewaarde[1,2], Roel Jonkers[1], and Roelien Bastiaanse[1,3]

1 Center for Language and Cognition, University of Groningen, r.a.ossewaarde@rug.nl, 2 HU University of Applied Science, Institute for ICT, Utrecht,

3 Center for Language and Brain, NRU Higher School of Economics, Moscow, Russia

Introduction Spontaneous speech is an important source of information for aphasia research. 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 expensive or places an undue burden on participants. The latter issue is an ethical consideration when working with participants that find speaking difficult, such as speakers with aphasia. So, how much speech data is enough to draw meaningful conclusions? 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 training?

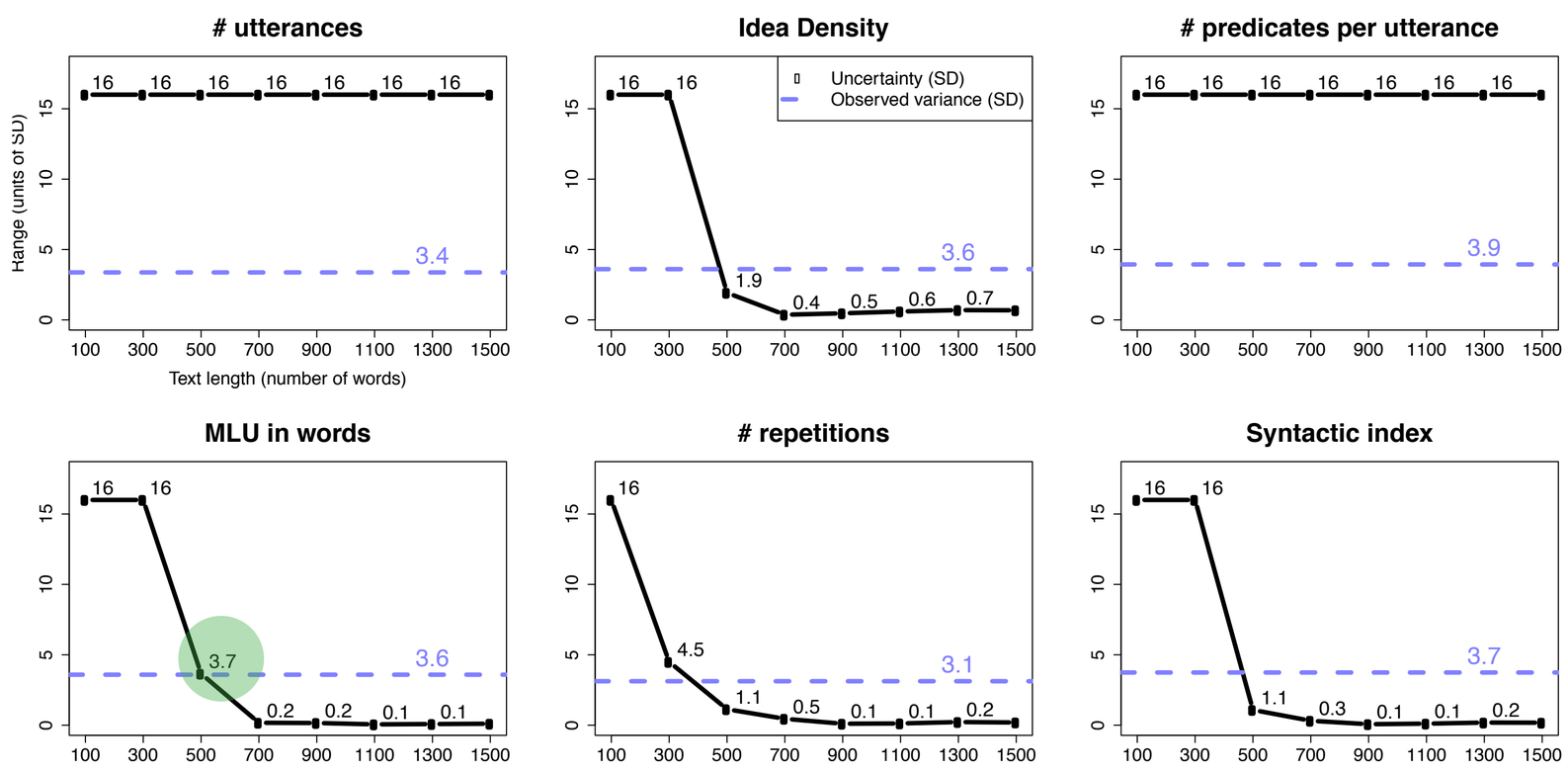
Methods & participants

We trained multiple regular regression models, each with data from the same corpus but truncated at different text length values. We then analyzed the uncertainty around the learned parameter values.

As training data, we use a convenience sample from a corpus of German spontaneous speech from non-brain-damaged speakers (NBDs; n=7) and individuals with a Alzheimer's Disease (AD; n=10).

The dependent variable is group membership (AD or control).

Fig. 2: Variation required to influence prediction, for various text lengths. Each point represents how far (max: 16 SD) the variable (ceteris paribus) must deviate from the mean to influence the final model prediction.



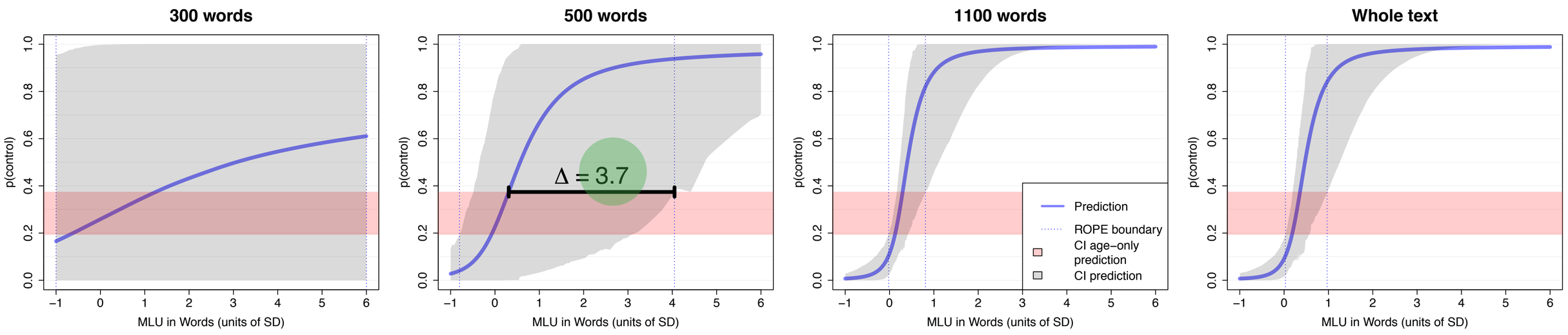


Fig. 1: The uncertainty around one variable when trained on texts with different lengths. When a predictor is significant, its value + uncertainty is more likely than in a random (age-only) model.

Parsing & statistical analysis

We used the German grammar of the Stanford Parser to measure text features, repeated for different text lengths. Parameters values and their uncertainty were estimated through Hamiltonian Monte Carlo simulation (RSTAN). We use Bayesian modeling because its interpretation of probability is a natural estimation of the relation between uncertainty and text length. The resulting model is compared to a model that includes only age as predictor to quantify how a model *with* linguistic variables fares better than a model *without* them.

Methods

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.

For each variable, we determined how much uncertainty is estimated at that text length. A variable is significant if it deviates more than its uncertainty from an age-only prediction (Fig. 1). When texts get longer, the uncertainty around significant variables decreases (Fig. 2).

Results

The uncertainty around individual variables decreases when longer narratives are analyzed.

A text length of **about 500** words is long enough for a linear model to distinguish significant variables. Fragments **longer than 700** words do not contribute more.

Our method is was applied to this particular corpus. Numbers would be different for different corpora, but the method can be applied just the same.