Predicting word 'predictability' in cloze completion, electroencephalographic and eye movement data

Chris Biemann[1], Steffen Remus[1] and Markus J. Hofmann[2]

[1] Language Technology Group, Comp. Sci. Dept., TU Darmstadt,
Hochschulstr. 10, 64289 Darmstadt, Germany
{biem,remus}@cs.tu-darmstadt.de
[2] General & Biological Psychology, Bergische Universität Wuppertal,
Max-Horkheimer Strasse 20, 42119 Wuppertal, Germany
mhofmann@uni-wuppertal.de

Abstract. Previous neurocognitive approaches to word predictability from sentence context in electroencephalographic (EEG) and eye movement (EM) data relied on cloze completion probability (CCP) data effortly collected from up to 100 human participants. Here we test whether two well-established techniques in computational linguistics can predict these data. Together with baseline predictors of word position and frequency, we found that n-gram language models but not topic models provide an approach to EEG and EM data that is not significantly inferior to the CCP-based predictability data. This is the case for the three corpora we used. Most strikingly, our models accounted for about half of the variance of the CCP-based predictability estimates, thus suggesting that it provides a computational framework to explain the predictability of a word from sentence context. This can help to generalize neurocognitive models to all possible novel word combinations.

1 Introduction

So far, manually collected cloze completion probabilities (CCPs) are typically used for quantifying a word's predictability from sentence context in neurocognitive psychology (Kutas and Hillyard, 1984; Reichle et al., 2003). Here we tackle the question whether the well-understood n-gram language models and Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) can account for CCPs, as well as whether they can provide an equally well-fitting approach to electroencephalographic (EEG) and eye movement (EM) measures, thus rendering time-consuming CCP procedures unnecessary.

CCPs have been traditionally used to account for N400 responses as an EEG signature of a word's contextual integration into sentence context (Dambacher et al., 2006; Kutas and Hillyard, 1984). Moreover, they were included as the quantification of the theoretical concept of predictability into models of eye movement control (Engbert et al., 2005; Reichle et al., 2003). However, because CCPs are effortly collected from samples of up to 100 participants (Kliegl et al., 2004), they provide a severe challenge to the ability of a model to be generalized across all novel stimuli (Hofmann and Jacobs, 2014), which also prevents their use in technical applications.

To quantify how well computational models of word recognition can account for human performance, Spieler and Balota (1997) proposed that a model should explain variance at the item-level, for instance naming latencies, averaged across a number of participants. Therefore, a predictor variable is fitted to the mean word naming latency y as a function of $y = f(x) = \sum a_n x_n + b + error$ for a number of n predictor variables x that are scaled by a slope factor a, an intercept of b, and an error term. The Pearson correlation coefficient r is calculated, and squared to determine the amount of explained variance r^2. Models with a larger number of n free parameters are more likely to (over-)fit error variance, and thus less free parameters are preferred (e.g., Hofmann and Jacobs, 2014).

While the best cognitive process models can account for 40-50% of variance in behavioral naming data (Perry et al., 2010), neurocognitive data are noisier. The only interactive activation model that gives an amount of explained variance in EEG data (Barber and Kutas, 2007; McClelland and Rumelhart, 1981) was Hofmann et al. (2008), who account for 12% of the N400 variance. Though models of eye movement control use item-level CCPs as predictor variables (Engbert et al., 2005; Reichle et al., 2003), they are rarely investigated in this field (Dambacher and Kliegl, 2007).

While using CCP-data increases the comparability of many studies, the creation of such information is expensive and they only exist for a few languages (Kliegl et al., 2004; Reichle et al., 2003). If it were possible to use (large) natural language corpora and derive the information leveraged from such resources automatically, this would considerably expedite the process of experimentation for under-resourced languages. Comparability would not be compromised when using standard corpora, such as available through Goldhahn et al. (2012) in many languages. However, it is not yet clear what kind of corpus is most appropriate for this enterprise, and whether there are differences in explaining human performance data.

2 Related Work

Taylor (1953) was the first to instruct participants to fill a cloze with an appropriate word. The percentage of participants that fill in the respective word serves as cloze completion probability. For instance, when exposed to the sentence fragment ”He mailed the letter without a ___”, 99% of the participants complete the cloze by ”stamp”, thus CCP equals 0.99 (Bloom and Fischler, 1980). Kliegl et al. (2004) logit-transformed CCPs to obtain $pred = ln(CCP/(1-CCP))$.

Event-related potentials are computed from human EEG data. For the case of the N400, words are often presented word-by-word, and the EEG waves are averaged across a number of participants relative to the event of word presentation. Because brain-electric potentials are labeled by their polarity and latency, the term N400 refers to a negative deflection around 400ms after the presentation of a target word.

After Kutas and Hillyard (1984) discovered the sensitivity of the N400 to cloze completion probabilities, they suggested that it reflects the semantic relationship between a word and the context in which it occurs. However, there are several other factors that determine the amplitude of the N400 (Kutas and Federmeier, 2011, for a review). For instance, Dambacher et al. (2006) found that word frequency (*freq*), the position of a word in a sentence (*pos*), as well as predictability does affect the N400.

While the eyes remain relatively still during fixations, readers make fitful eye movements called saccades (Radach et al., 2012). When successfully recognizing a word in a stream of forward eye movements, no second saccade to or within the word is required. The time the eyes remain on that word is called single-fixation duration (SFD), which shows a strong correlation to word predictability from sentence context (e.g., Engbert et al., 2005).

3 Methodology

3.1 Human Performance Measures

This study proposes that language models can be benchmarked by item-level performance on three data sets that are openly available in online databases. Predictability was taken from the Potsdam Sentence Corpus 1, first published by Kliegl et al. (2004). The 144 sentences consist of 1138 tokens, available in Appendix A of Dambacher (2009), and the logit-transformed CCP measures of word predictability were retrieved from Ralf Engbert's homepage[1] (Engbert et al., 2005). For instance, in the sentence "Manchmal sagen Opfer vor Gericht nicht die volle Wahrheit" [Before the court, victims tell not always the truth.], the last word has a CCP of 1. N400 amplitudes were taken from the 343 open-class words published in Dambacher and Kliegl (2007). These are available from the Potsdam Mind Research Repository[2]. The EEG data published there are based on a previous study (Dambacher et al., 2006, for method details). The voltage of ten centroparietal electrodes was averaged across 48 artifact-free participants from 300 to 500ms after word presentation for quantifying the N400. SFD are based on the same 343 words from Dambacher and Kliegl (2007), available from the same source URL. Data were included when this word was only fixated for one time, and these SFDs ranged from 50 to 750ms. The SFD was averaged across up to 125 German native speakers (Dambacher and Kliegl, 2007).

3.2 N-gram Language and LDA Topic Models

Language models are based on a probabilistic model of language. The resulting probabilities can be used to pick the most fluent of several alternatives e.g. in machine translation or speech recognition. Word **n-gram models** are defined by a Markov chain of order $n-1$, where the probability of the following word only depends on previous $n-1$ words. The probability distribution of the vocabulary, given a history of $n-1$ words, is estimated based on n-gram counts from (large) natural language corpora. There exist a range of n-gram language models (see for example Chapter 3 in Manning and Schütze, 1999). Here, we use a Kneser and Ney (1995) 5-gram mod-

[1] http://mbd.unipotsdam.de/EngbertLab/Software.html
[2] http://read.psych.unipotsdam.de

el[3]. For each word in the sequence, the language model computes a probability $p \in]0;1[$. We use the logarithm *log(p)* of this probability as predictor. We used all words in their full form, i.e. did not filter for specific word classes and did not perform lemmatization. N-gram language models are known to model local syntactic structure very well. Since only n-gram models use the most recent history for predicting the next token, they fail to account for long-range phenomena and semantic coherence, cf. (Biemann et al., 2012).

Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) are generative probabilistic models representing documents as a mixture of a fixed number of N topics, which are defined as unigram probability distributions over the vocabulary. Through a sampling process like Gibbs sampling, topic distributions are inferred. Words frequently co-occurring in the same documents receive a high probability in the same topics. When sampling the topic distribution for a sequence of text, each word is randomly assigned to a topic according to the document-topic distribution and the topic-word distribution. We use Phan and Nguyen's (2007) GibbsLDA implementation for training an LDA model with 200 topics (default values for $\alpha = 0.25$ and $\beta = 0.001$) on a background corpus. Words occurring in too many documents (a.k.a. stopwords) were removed from the LDA vocabulary. Then, we repeatedly sample the topic assignments (cf. Riedl and Biemann, 2012) on the input sentence and retain the most frequently assigned three topics per word. As predictor for the current open class word in the sequence, we count the number of previous open class words in the sequence, which have at least one topic in common with the current word. Intuitively, this measure should capture the amount of semantic coherence with the previous words in the sequence. I.e. for a sequence like "The dwarf was avoiding the ____", we'd expect a score of 1 for "elves" for their topical similarity to "dwarf" (provided that there is sufficient support of dwarves and elves in the background corpus), whereas we expect a score of 0 for "rain". Parameters of this procedure were determined in preliminary experiments. We hypothesized that topic models account for the semantic aspects missing in n-gram models. While Bayesian topic models are probably the most widespread approach to semantics in psychology (e.g., Griffiths et al., 2007), latent semantic analysis (LSA) is not applicable in our setting (Landauer and Dumais, 1997): we use the capability of LDA to account for yet unseen documents, whereas LSA assumes a fixed vocabulary and document space at model construction time. In further experiments, we also used collocation statistics to predict semantically expected items, but we obtained no correlation with human data.

4 Experiment Setup

Engbert et al. (2005)'s data are organized in 144 short German sentences with an average length of 7.9 tokens, and provide features, such as *freq* as corpus frequency in occurrences per million (Baayen et al., 1995), *pos*, and *pred*. We test whether two corpus-based predictors can account for predictability, and compare the capability of both approaches in accounting for EEG and EM data. For training n-gram and topic

[3] https://code.google.com/p/berkeleylm/

models, we used three different corpora differing in size and covering different aspects of language. Further, the units for computing topic models differ in size.

NEWS: A large corpus of German online newswire from 2009 as collected by LCC (Goldhahn et al., 2012) of 3.4 million documents / 30 million sentences / 540 million tokens. This corpus is not balanced, i.e. important events in the news are covered better than other themes. The topic model was trained on the article level.

WIKI: A recent German Wikipedia dump of 114,000 articles / 7.7 million sentences / 180 million tokens. This corpus is rather balanced, as concepts or entities are described in a single article each, independent of their popularity, and spans all sorts of topics. The topic model was trained on the article level.

SUB German subtitles from a recent dump of opensubtitles.org, containing 7420 movies / 7.3 million utterances / 54 million tokens. While this corpus is much smaller than the others, it is closer to a colloquial use of language. Brysbaert et al. (2011) showed that word frequency measures of subtitles provide numerically greater correlations with word recognition speed than larger corpora of written language. The topic model was trained on the movie level.

Pearson's product-moment correlation coefficient was calculated (e.g. Coolican, 2010, p. 293), and squared for the N = 1138 predictability scores (Engbert et al., 2005) or N = 343 N400 amplitudes or SFD (Dambacher and Kliegl, 2007). To address overfitting, we randomly split the material in two halves, and test how much variance can be reproducibly predicted on two subsets of 569 items. For N400 amplitude and SFD, we used the full set, because one half was too small for reproducible predictions.

5 Results

5.1 Predictability results

In the first series of results, we examine the correlation of manually obtained predictability with corpus-based methods. High correlations would indicate that predictability could be replaced by automatic methods. As a set of baseline predictors, we use *pos* and *freq*, which explains 0.243 / 0.288 of the variance for the first respectively the second half of the dataset. We report results in Table 1 for all single corpus-based predictors alone and in combination with the baseline, all combinations of the baseline with n-grams and topics from the same corpus.

predictors	NEWS	WIKI	SUB
n-gram alone	.262/.294	.226/.253	.268/.272
topic alone	.024/.037	.029/.022	.014/.012
base+n-gram	.462/.490	**.462/.490**	**.448/.459**
base+topic	.252/.307	.254/.296	.244/.289
base+both	**.481/.516**	.445/.473	**.449/.461**

Table 1. r^2 explained variance of predictability, given for two folds of the data set, for various combinations of baseline and corpus-based predictors.

It is apparent that the n-gram predictor alone reaches r^2 levels comparable to the baseline, whereas the topic model alone explains hardly any variance. Combining the baseline with the n-gram predictor achieves the best fitting to predictability for the WIKI and SUB corpora. Combining the baseline with topics shows small improvements for NEWS and WIKI (see Figure 1).

The best overall performance based on a single corpus is achieved with combining the baseline with n-grams and topics from the NEWS corpus. This confirms a generally accepted hypothesis that larger training data trumps smaller, more focused training data, see e.g. (Banko and Brill, 2001) and others. We also fitted a model based on all corpus-based predictors from all corpora, which achieved the overall highest r^2= 0.532 / 0.547 . From these experiments it becomes clear that predictability can largely be explained by a combination positional and frequency features combined with a word n-gram language model. Different corpora capture slightly different aspects of predictability, which is reflected by the improvements when combining predictors from all three corpora. The topic model-based predictor only shows a negligible influence.

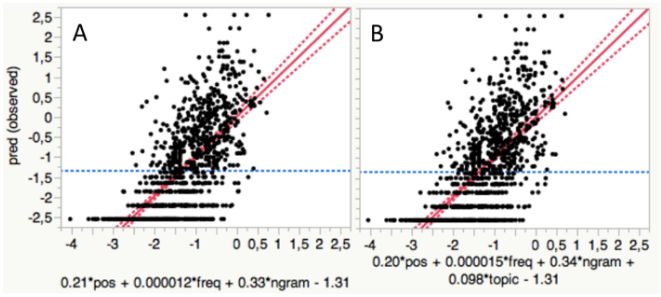

Fig. 1. Prediction models exemplified for the NEWS corpus in the x-axes and the N = 1138 predictability scores on the y-axes. A) shows the prediction by baseline + n-gram (r^2=0.475), and in B) a topic-predictor was added (r^2=0.481). Fisher's r-to-z test revealed that there is no significant difference in explained variance (P=0.82)

5.2 N400 and SFD results

For modeling **N400**, we have even more combinations at our disposal since we can combine the baseline with predictability as given in the dataset, with corpus-based measures, and with both. We evaluate on all 343 data points for N400 amplitude fitting. Without using corpus-based predictors, the baseline predicts a mere 0.032 of variance, predictability alone explains 0.192 of variance, and their combination explains 0.193 – i.e. the baseline is almost entirely subsumed by predictability.

Fig. 2 lists the results for N400 amplitude modeling with corpus-based predictors. Again, the n-gram model is the best corpus-based predictor, and fares best when trained on the NEWS corpus, confirming the result that corpus size is the major factor for n-gram model quality. For the N400 experiments, the difference between the larger corpora (NEWS, WIKI) and the smaller corpus (SUB) is more pronounced. Again, the topic predictor fails to show a major influence for explaining N400 amplitude

variance. The best combination without predictability, with a score of r^2 = 0.182, comes close to the performance of predictability alone.

predictors	NEWS	WIKI	SUB
n-gram alone	0.141	0.140	0.126
topic alone	0.022	0.021	0.006*
n-gram+topic	0.170	0.166	0.131
base+n-gram	0.161	0.153	0.135
base+topic	0.051	0.050	0.036
bas+n-gram+topic	**0.182**	0.172	0.137
base+pred+n-gram	**0.223**	**0.226**	0.206
base+pred+topic	0.194	0.193	0.193
base+pred+both	**0.228**	**0.229**	0.206

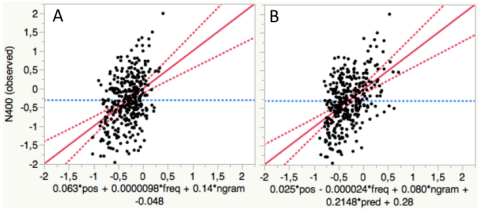

Fig. 2. Left: r^2 explained variance of N400 amplitude, for various combinations of baseline, predictability and corpus-based predictors. * marks statistically independent predictors of N400 ($p > 0.05$). Right: Two prediction models exemplified for the NEWS corpus in the x-axes and the N = 343 N400 amplitudes on the y-axes. A) shows the prediction by baseline + n-gram, and in B) predictability was added. Fisher's r-to-z test revealed that there is no significant difference in explained variance (P=0.25)

The experiments with predictability as an additional predictor confirm the results from the previous section: n-grams + baseline and predictability capture slightly different aspects of human reading performance, thus their combination explains about 3% more variance than predictability alone. This difference, however, is not statistically reliable (see Figure 2). Differences between the two large corpora are negligible, and so is the influence of the topic-based predictor.

Finally, we examine the corpus-based predictors for modeling the mean **single fixations duration** for 343 words. For this target, the *pos+freq* baseline explains r^2 = 0.021, whereas predictability, alone or combined with the baseline, explains r^2 = 0.184.

predictors	NEWS	WIKI	SUB
n-gram alone	0.225	0.140	0.126
topic alone	0.006*	0.006*	0.006*
n-gram+topic	0.231	0.223	0.226
base+n-gram	**0.239**	0.226	0.226
base+topic	0.023	0.024	0.029
bas+n-gram+topic	**0.242**	0.230	0.229
base+pred+n-gram	**0.273**	**0.274**	0.258
base+pred+topic	0.188	0.184	0.184
base+pred+both	**0.273**	**0.274**	0.259

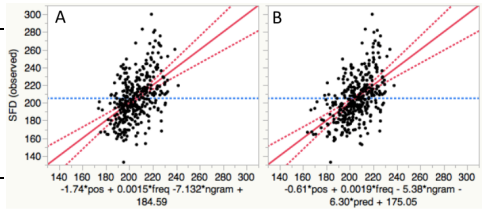

Fig. 3. Left: r^2 explained variance of single-fixation duration, for various combinations of baseline, predictability and corpus-based predictors. * marks statistically independent predictors of SFD ($P > 0.05$). Right: Two prediction models exemplified for the NEWS corpus in the x-axes and the N = 343 SFD on the y-axes. A) shows the prediction by baseline + n-gram, and in B) predictability was added. Fisher's r-to-z test revealed that there is no significant difference in explained variance (P=0.56)

The experiments confirm the utility of n-gram models in accounting for eye movement data. Adding predictability did not lead to a significant increase of vari-

ance explained (see Fig. 3). In addition, the n-gram model alone explains more variance than predictability – however, the difference is not significant.

For SFD, corpus size does not seem to be a major influencing factor, as results are comparable across corpora, however with the largest corpus (NEWS) still yielding the best modeling results overall in absence of the predictability predictor. For SFD, topic models seem entirely uncorrelated.

And again, the experiments confirm that n-gram models and predictability capture similar, but slightly different aspects, since their combination yields another improvement, explaining $r^2 = 0.273$ overall.

6 Conclusion

We have examined the utility of two corpus-based predictors to account for word predictability from sentence context, as well as the EEG signals and EM-based reading performance elicited by it. Our hypothesis was that word n-gram models and topic models would account for the predictability of a token, given the preceding tokens in the sentence, as perceived by humans. Our hypothesis was at least partially confirmed: n-gram models, sometimes in combination with a frequency-based and positional baseline, are highly correlated with human predictability scores and in fact explain variance of human reading performance to an extent comparable to predictability – slightly less on N400 but slightly more on SFD.

Topic models on the other hand, at least in the particular way we used them here, failed to show a major influence on modeling human reading performance. This might be related to the fact that the sentence scope in the data set is rather short so that most "priming" effects can already be captured by our 5-gram model – topic models usually perform well on the level of documents, not single sentences.

Can we now safely replace human predictability scores with n-gram statistics? Given the high correlation between predictability and the combination of n-grams with frequency and positional information, and given that n-gram-based predictors achieve similar levels of explained variance than predictability, the answer seems to be positive. However, though our corpus-based approaches explain most of the variance that by manually collected CCP scores also account for, adding predictability always accounts for more variance – though this difference is not significant (see Figures 1-3). It is yet an open question, whether additional corpus-based predictors, be it topic models or something else, could entirely explain the prediction power of human CCP data for tasks like N400 amplitude and SFD modeling.

While n-gram models together with word frequency and position captured about half of the predictability variance, and most of the N400 and SFD variance elicited by it, we propose that it can be used to replace tediously collected CCPs. This not only saves a lot of pre-experimental work, but it also opens the possibility to apply (neuro-) cognitive models in technical applications. For instance, n-gram models can be used to generalize computational models of eye movement control to novel sentences (Engbert et al., 2005; Reichle et al., 2003).

In the end, this will also improve our understanding of the cognitive processes underlying EM and EEG measures. While both of these are not as well understood as human CCP performance, predictability provided a great step towards understanding

the determinants of neurocognitive prediction processes. If we can compute the determinants of N400 and SFDs from a corpus of sentences, however, we can computationally define these cognitive processes rather than using a better-understood performance (CCP) to account for other human performance (N400, SFD).

Baayen (2010) proposed word frequency to be a collector variable often subsuming other highly correlated variables. We found that adding n-grams to the baseline of pos and freq doubled the explained variance in CCP-based predictability scores. This suggests that the sentence level can unfold the cognitive processes previously ascribed to word frequency. The doubling of explained variance suggests still unexploited sources of human variance to be explained by neurocognitive simulation models, which quantify the contextual constraints imposed by position-sensitive predictions of a sentence's words (e.g. Hofmann & Jacobs, 2014; Kutas & Federmeier, 2011).

Much as for computational models of word recognition, the amount of explained item-level variance can serve as a benchmark for language models. Such a common benchmark facilitates the comparison of differential computational models. Thus, for instance, we would not only know that Frank et al. (2013)'s novel language model can account for the N400, but the common benchmark of explained variance could be easily compared to any novel approach – for instance by assessing whether one measure is significantly better than another one for the purpose of modeling.

Acknowledgments: The "Deutsche Forschungsgemeinschaft" (MJH; HO 5139/2-1), the German Institute for Educational Research in the Knowledge Discovery in Scientific Literature (SR) program and the LOEWE center for Digital Humanities (CB) supported this work.

References

H. R. Baayen, R. Piepenbrock, and L. Gulikers (1995) The CELEX Lexical Database. Release 2 (CD-ROM). LDC, University of Pennsylvania, Philadelphia.

H.R. Baayen (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3):436-461.

M. Banko and E. Brill (2001) Scaling to very very large corpora for natural language disambiguation. *Proc. ACL '01*, pp. 26–33, Toulouse, France.

H. A. Barber and M. Kutas (2007) Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Res. Rev.*, 53(1):98–123.

C. Biemann, S. Roos, K. Weihe (2012) Quantifying semantics using complex network analysis. *Proc. COLING 2012*, pp. 263–278, Mumbai, India.

D. M. Blei, A. Y. Ng, M. I. Jordan (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

P. A. Bloom and I. Fischler (1980) Completion norms for 329 sentence contexts. *Memory & cognition*, 8(6):631–642.

M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl (2011) A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental psychology*, 58:412–424.

H. Coolican (2010) Research Methods and Statistics in Psychology. *Hodder & Stoughton*.

M. Dambacher and R. Kliegl (2007) Synchronizing Timelines: Relations between fixation durations and N400 amplitudes during sentence reading. *Brain research*, 1155:147–162.

M. Dambacher, R. Kliegl, M. J. Hofmann, A. M. Jacobs. (2006) Frequency and predictability

effects on event-related potentials during reading. *Brain research*, 1084(1):89–103.

M. Dambacher. 2009. Bottom-up and top-down processes in reading. *Universitätsverlag Potsdam*, Potsdam.

R. Engbert, A. Nuthmann, E. M. Richter, R. Kliegl (2005) SWIFT: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777–813.

S. L. Frank, G. Galli, and G. Vigliocco (2013) Word surprisal predicts N400 amplitude during reading. *Proc. ACL-2013,* pp. 878–883, Sofia, Bulgaria.

D. Goldhahn, T. Eckart, U. Quasthoff (2012) Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. *Proc. LREC 2012*, Istanbul, Turkey.

T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum (2007) Topics in Semantic Representation. *Psychological review*, 114(2):211–244.

M. J. Hofmann and A. M. Jacobs (2014) Interactive activation and competition models and semantic context: From behavioral to brain data. *Neuroscience and biobehav. rev.s*, 46:85–104.

M. J. Hofmann, S. Tamm, M. M. Braun, M. Dambacher, A. Hahne, and A. M. Jacobs (2008) Conflict monitoring engages the mediofrontal cortex during nonword processing. *Neuroreport*, 19(1):25–9.

R. Kliegl, E. Grabner, M. Rolfs, and R. Engbert (2004) Length, frequency, and predictability effects of words on eye movements in reading. *Europ. Journal of Cog. Psy.*, 16(12):262–284.

R. Kneser and H. Ney (1995) Improved backing-off for m-gram language modeling. *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, pp. 181–184, Detroit, Michigan.

M. Kutas and K. D. Federmeier (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Ann. Rev. of Psychology*, 62:621–47.

M. Kutas and S. A. Hillyard (1984) Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–3.

T. K. Landauer and S. T. Dumais (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

C. D. Manning and H. Schütze (1999) Foundations of Statistical Natural Language Processing. *MIT Press*, Cambridge, MA, USA.

J. L. McClelland and D. E. Rumelhart (1981) An Interactive Activation Model of Context Effects in Letter Perception: Part 1. *Psychological Review*, 5:375–407.

C. Perry, J. C. Ziegler, and M. Zorzi (2010) Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, 61(2):106–51.

X-H. Phan and C-T. Nguyen (2007) GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). http://jgibblda.sourceforge.net/.

R. Radach, T. Günther, and L. Huestegge (2012) Blickbewegungen beim Lesen, Leseentwicklung und Legasthenie. *Lernen und Lernstoerungen*, 1(3):185–204.

E. D. Reichle, K. Rayner, and A. Pollatsek (2003) The E-Z reader model of eye-movement control in reading: comparisons to other models. *The Behavioral and brain sciences*, 26(4):445–76; discussion 477–526.

M. Riedl and C. Biemann (2012) Sweeping through the topic space: Bad luck? roll again! *Proc. ROBUS-UNSUP 2012*, Avignon, France.

D. H. Spieler and D. A. Balota (1997) Bringing Computational Models of Word Naming Down to the Item Level. *Psychological Science*, 8(6):411–416

W. L. Taylor (1953) "Cloze" procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415.