Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements

Chapter · March 2017

CITATIONS
0

READS
29

3 authors, including:

Markus J. Hofmann
Bergische Universität Wuppertal
48 PUBLICATIONS **1,070** CITATIONS
SEE PROFILE

Chris Biemann
University of Hamburg
117 PUBLICATIONS **1,134** CITATIONS
SEE PROFILE

1

Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements

Previous neurocognitive approaches to word predictability from sentence context in electroencephalographic (EEG) and eye movement (EM) data relied on cloze completion probability (CCP) data effortfully collected from up to 100 human participants. Here we test whether three well-established language models can predict these data. Together with baseline predictors of word frequency and position in sentence, we found that the syntactic and short-range semantic processes of n-gram language models and recurrent neural networks (RNN) perform about equally well when directly accounting CCP, EEG and EM data. In contrast, a low amount of variance explained by a topic model suggests that there is no strong impact on the CCP and the N400 component of the EEG, at least in our Potsdam Sentence Corpus dataset. For the single-fixation durations of the EM data, however, topic models accounted for more variance, suggesting that long-range semantics may play a greater role in this earlier neurocognitive process. Though the language models were not significantly inferior to CCP in accounting for these EEG and EM data, CCP always provided a descriptive increase in explained variance for the three corpora we used. However, n-gram and RNN models can account for about half of the variance of the CCP-based predictability estimates, and the largest part of the variance that CCPs explain in EEG and EM data. Thus, our approaches may help to generalize neurocognitive models to all possible novel word combinations, and we propose to use the same benchmarks for language models than for models of visual word recognition.

Chapter written by Markus J. HOFMANN, Chris BIEMANN, and Steffen REMUS

1.1. Introduction

In neurocognitive psychology, manually collected cloze completion probabilities (CCPs) are the standard approach to quantifying a word's predictability from sentence context [KLI 04; KUT 84; REI 03]. Here we test a series of language models in accounting for CCPs, as well as the data they typically account for, i.e. electroencephalographic (EEG) and eye movement (EM) data. With this, we hope to render time-consuming CCP procedures unnecessary. We test a statistical n-gram language model [KNE 95], a Latent Dirichlet Allocation (LDA) topic model [BLE 03], as well as a recurrent neural network (RNN) language model [BEN 03; ELL 90] for correlation with the neurocognitive data.

CCPs have been traditionally used to account for N400 responses as an EEG signature of a word's contextual integration into sentence context [DAM 06; KUT 84]. Moreover, they were used to quantify the concept of word predictability from sentence context in models of eye movement control [ENG 05; REI 03]. However, because CCPs are effortfully collected from samples of up to 100 participants [KLI 04], they provide a severe challenge to the ability of a model to be generalized across all novel stimuli [HOF 14], which also prevents their ubiquitous use in technical applications.

To quantify how well computational models of word recognition can account for human performance, Spieler and Balota [SPI 97] 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 fit 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., [HOF 14]).

While the best cognitive process models can account for 40-50% of variance in behavioral naming data [PER 10], neurocognitive data are noisier. The only interactive activation model that gives an amount of explained variance in EEG data [BAR 07; MCC 81] was Hofmann et al. [HOF 08], who account for 12% of the N400 variance. Though models of eye movement control use item-level CCPs as predictor variables [ENG 05; REI 03], computational models of eye movement control have hardly been benchmarked at the item-level, to our knowledge [DAM 07].

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 [KLI 04; REI 03]. 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. [GOL 12] 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.

1.2. Related Work

Taylor [TAY 53] 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 [BLO 80]. Kliegl et al. [KLI 04] 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 [KUT 84] 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 [KUT 11]. For instance, Dambacher et al. [DAM 06] 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 [RAD 12]. 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., [ENG 05]).

1.3. Methodology

1.3.1 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. [KLI 04]. The 144 sentences consist of 1138 tokens, available in Appendix A of [DAM 09], and the logit-transformed CCP measures of word predictability were retrieved from Ralf Engbert's homepage[1] [ENG 05]. 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 [DAM 07]. These are available from the Potsdam Mind Research Repository[2]. The EEG data published there are based on a previous study (see [DAM 06] 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 [DAM 07], 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 [DAM 07].

1.3.2 Three Flavors of Language Models

Language models are based on a probabilistic description of language phenomena. Probabilities are 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. In statistical models, 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 [MAN 99], which are differentiated by the way they handle unseen events and perform probability smoothing). Here, we use a Kneser-Ney [KNE 95] 5-gram model[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 mod-

[1] http://mbd.unipotsdam.de/EngbertLab/Software.html
[2] http://read.psych.unipotsdam.de
[3] https://code.google.com/p/berkeleylm/

els 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. [BIE 12].

Latent Dirichlet Allocation (LDA) topic models [BLE 03] 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 [PHA 07] 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) or too few documents (mistyped or rare words) were removed from the LDA vocabulary. Then, retain the per document topic distribution $p(z|d)$ and the per topic word distribution $p(w|z)$, where z is the latent variable representing the topic, d refers to a full document during training—during testing d refers to the history of the current sentence—and w is a word. In contrast to our earlier approach using only the top three topics [BIE 15], we here computed the probability of the current word w given its history d as a mixture of its topical components $p(w|d) = p(w|z)p(z|d)$. We hypothesize that topic models account for some long-range semantic aspects missing in n-gram models. While Bayesian topic models are probably the most widespread approach to semantics in psychology (e.g., [GRI 07]), latent semantic analysis (LSA) is not applicable in our setting [LAN 97]: we use the capability of LDA to account for yet unseen documents, whereas LSA assumes a fixed vocabulary and it is not trivial to fold new documents into LSA's fixed document space.

While Jeff Elman's [ELM 90] seminal work suggested already early that semantic and also syntactic structure automatically emerges from a set of simple recurrent units, such an approach has received little attention in language modelling for a long time, but is currently in focus of many computational studies. In brief, such **Neural Network Language Models** are based on the optimization the probability of the occurrence of a word given its history using neural units linking back to themselves, much as the neurons in the CA3 region of the human hippocampus [MAR 71, NOR 03]. The task of language modelling using neural networks was first introduced by Bengio et al. [BEN 03] and received at that point only little attention because of computational challenges regarding space and time complexity. Due to recent advancement in the field of neural networks—for an overview see [MIK 12]—neural language models gained a more popularity, particularly because of so-called neural word embeddings as a side product. The language model implementation we use in

this work is a recurrent neural network architecture[4] similarly to the one used by Mikolov's Word2Vec[5] toolkit [MIK 13]. We trained a model with 400 hidden layers and hierarchical softmax. For testing we used the complete history of a sentence up to the current word.

1.4. Experiment Setup

Engbert et al.'s [ENG 05] 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 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 [GOL 12] 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. [BRY 11] 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. [COO 10], p. 293), and squared for the N = 1138 predictability scores [ENG 05] or N = 343 N400 amplitudes or SFD [DAM 07]. 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. The correlations between

[4] FasterRNN: https://github.com/yandex/faster-rnnlm
[5] Word2Vec: https://code.google.com/archive/p/word2vec/

all predictor variables can be examined in Table 1.1. We observe very high correlations between the n-gram and the RNN predictions within and across corpora. The correlations involving topic-based predictions are smaller, supporting our hypothesis that they reflect a somewhat different neurocognitive process.

		1.	2.	3.	4.	5.	6.	7.	8.	9.
NEWS	1. n-gram		0.65	0.87	0.87	0.56	0.84	0.83	0.59	0.80
	2. topic	0.65		0.68	0.66	0.78	0.70	0.61	0.77	0.61
	3. neural	0.87	0.68		0.84	0.59	0.88	0.77	0.62	0.79
WIKI	4. n-gram	0.87	0.66	0.84		0.61	0.90	0.79	0.59	0.78
	5. topic	0.56	0.78	0.59	0.61		0.65	0.55	0.75	0.55
	6. neural	0.84	0.70	0.88	0.90	0.65		0.76	0.64	0.79
SUB	7. n-gram	0.83	0.61	0.77	0.79	0.55	0.76		0.61	0.85
	8. topic	0.59	0.77	0.62	0.59	0.75	0.64	0.61		0.61
	9. neural	0.80	0.61	0.79	0.78	0.55	0.79	0.85	0.61	

Table 1.1. *Correlations between the language model predictors*

1.5. Results

1.5.1 Predictability results

In the first series of results, we examine the prediction of manually obtained CCP-derived predictability with corpus-based methods. A large amount of explained variance 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 1.2 for all single corpus-based predictors alone and in combination with the baseline, all combinations of the baseline with n-gram, topics, and neural models from the same corpus.

Predictors	NEWS	WIKI	SUB
n-gram	.262/.294	.226/.253	.268/.272
topic	.063/.061	.042/.040	.040/.034
neural	.229/.226	.211/.226	.255/.219
base+n-gram	**.462/.490**	.423/.458	.448/.459
base+topic	.348/.375	.333/.357	.325/.355
base+neural	**.434/.441**	.418/.433	**.447/.418**
base+n-gram+topic	.462/.493	.427/.464	.447/.458
base+n-gram+neural	.466/.492	.431/.461	.467/.461
base+neural+topic	.438/.445	.421/.436	.446/.423
base+n-gram+topic+neural	.466/.493	.433/.465	.467/.460

Table 1.2. *r^2 explained variance of predictability, given for two halves of the data set, for various combinations of baseline and corpus-based predictors*

It is apparent that the n-gram scores best, but also the neural model alone reach r^2 levels that approach the baseline. In contrast, much as our earlier top-three topics approach [BIE 15], the mixture of all topics explains only a relatively low amount of variance. Combining the baseline with the n-gram predictor already reaches a level very close to the combination of all predictors, thus it may provide the best compromise between parsimony and explained variance. Again, this model performance is closely followed by the recurrent neural network (see Figure 1.1).

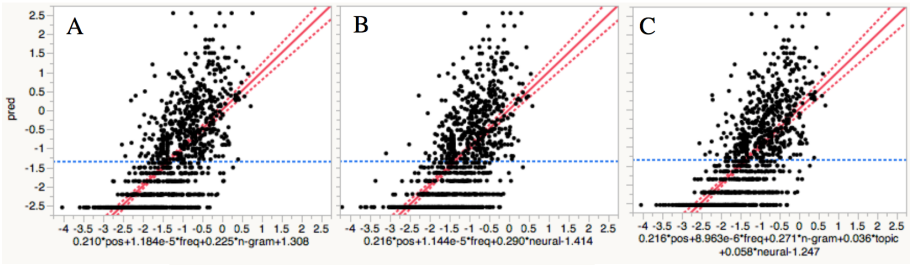

Figure 1.1. *Prediction models exemplified for the NEWS corpus in the x-axes and the N = 1138 predictability scores on the y-axes. A) shows the prediction by baseline + n-gram (r^2=0.475), B) a recurrent neural network (r^2=0.437), and C) a model containing all predictors (r^2=0.478). The three pairwise Fisher's r-to-z tests revealed no significant differences in explained variance (Ps>0.18)*

We also fitted a model based on all corpus-based predictors from all corpora, which achieved the overall highest r^2= 0.490/.507. In sum, it becomes clear that about half of the empirical predictability variance can be explained by a combination positional and frequency features combined with either a word n-gram language model, or a recurrent neural network.

1.5.2. N400 amplitude results

For modeling N400, we have even more combinations at our disposal since we can combine corpus-based measures with the baseline, the predictability performance, 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 CCP-based predictability. As can be seen in Table 1.3, this is a score that is not yet reached by the language models, even when combining all of them.

Predictors	NEWS	WIKI	SUB
n-gram	.141	.140	.126
topic	.039	**.055**	**.**025
neural	.108	.098	.100
base+n-gram	**.161**	.153	.135
base+topic	.063	.079	.055
base+neural	.133	.116	.114
base+n-gram+topic	.161	.158	.132
base+n-gram+neural	.167	.153	.141
base+neural+topic	.133	.123	.112
base+n-gram+topic+neural	.**167**	.158	.137
base+n-gram+pred	.**223**	.226	.206
base+topic+pred	.193	.204	.191
base+neural+pred	.221	.212	.206
base+n-gram+topic+pred	.225	.228	.203
base+n-gram+neural+pred	.228	.226	.209
base+neural+topic+pred	.224	.215	.203
base+n-gram+topic+neural+pred	.**232**	.228	.206

Table 1.3. *r^2 explained variance of the N400 for various combinations of the corpus-based predictors, in combination with the baseline, and with the empirical predictability.*

When comparing the performance of the computationally defined predictors, a similar picture as with the prediction of the empirical predictability emerges. The n-gram model scores best, particularly at the larger NEWS and WIKI corpora. This confirms a generally accepted hypothesis that larger training data trumps smaller, more focused training data, see e.g. [BAN 01] and others. The n-gram model is however immediately followed by the neural model, and again, the topic predictor provides the poorest performance in explaining N400 amplitude variance, which suggests that the N400 does not reflect long-range semantic processes. The best combination without predictability, with a score of r^2 = 0.167 approaches the performance of the predictability and baseline (cf. Figure 1.2).

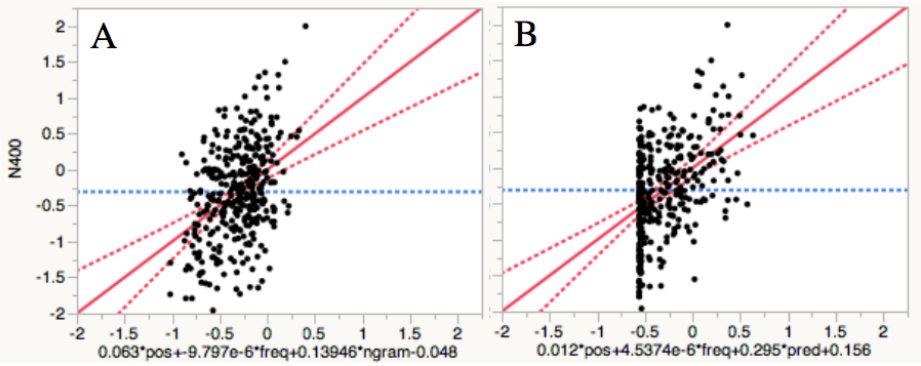

Figure 1.2. *Prediction models exemplified for the NEWS corpus in the x-axes and the N = 334 mean N400 amplitudes on the y-axes. A) shows the prediction by baseline + n-gram (r^2=.161), and B) shows a standard approach to N400 data, consisting of the baseline of position and frequency, as well as the empirical predictability (r^2=.193; e.g. Dambacher, Kliegl, Hofmann, & Jacobs, 2006). Fisher's r-to-z tests revealed no significant differences in explained variance (P = 0.55)*

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 up to 6% more net variance than predictability alone.

1.5.3 Single-Fixation Duration (SFD) results

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	.225	.140	.126
topic	.135	.140	.100
neural	**.**242	.190	.**272**
base+n-gram	.239	.226	.226
base+topic	.152	.154	.127
base+neural	.265	.204	.**284**
base+n-gram+topic	.260	.262	.246
base+n-gram+neural	.287	.238	.297
base+neural+topic	.279	.235	.298
base+n-gram+topic+neural	**.**295	.265	.**307**
base+n-gram+pred	.273	.274	.258
base+topic+pred	.235	.250	.229
base+neural+pred	.314	.267	.320
base+n-gram+topic+pred	.297	.301	.275
base+n-gram+neural+pred	.319	.283	.322
base+neural+topic+pred	.319	.289	.329
base+n-gram+topic+neural+pred	.323	.304	.**330**

Table. 1.4. *Explained variance of the single-fixation durations, for various combinations of baseline, predictability and corpus-based predictors*

The experiments confirm the utility of n-gram models in accounting for eye movement data. The n-gram model alone explains even more variance than predictability – however, the difference is not significant ($P > .46$).

In contrast to the previous approaches to predictability and N400 amplitudes, however, the recurrent neural network outperformed the n-gram model at a descriptive level, as it accounted for up to 3% more of the variance than the n-gram model. This performance was not reached at the largest NEWS corpus, but at the smaller SUB corpus. This suggests that – for SFD data – the dimension reduction seems to compensate for the larger amount of the noise in the smaller training data set, cf. [BUL 07; GAM 16*; HOF 14]. Therefore, the neural model may provide a better fit for such early neurocognitive processes when it is trained by colloquial language [BRY 11].

The topics model seems to have a stronger impact on SFDs than on the other neurocognitive benchmark variables, suggesting a greater influence of long-range semantics on SFDs than on predictability or the N400. Taken together, these findings point towards SFDs to reflect different cognitive processes than the N400 (cf. [DAM 07]).

Last but not least, though again adding predictability increased the total amount of explained variance by 2%, the language models did an excellent job in accounting for SFD data. When taking all language-model based predictors together, this accounts for significantly more variance than the standard model using predictability (see Fig. 1.3).

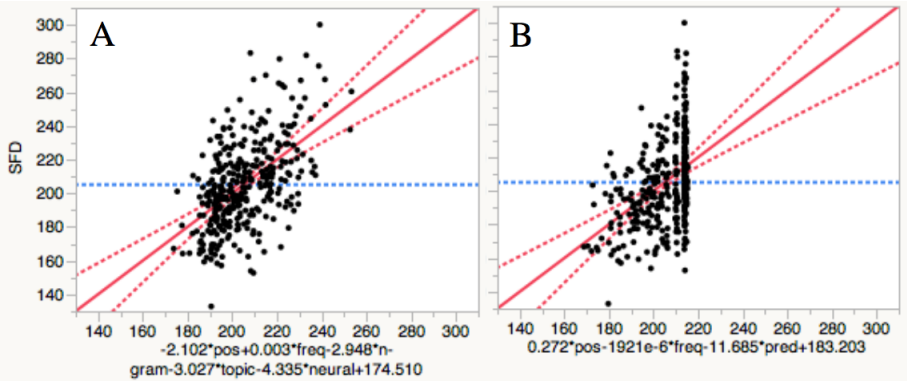

Figure 1.3. *Prediction models exemplified for the SUB corpus in the x-axes and the N = 334 mean SFD scores on the y-axes. A) shows the prediction by baseline + all three language model (r^2=.295), and B) shows a standard approach to SFD data, using the baseline and predictability as predictors of SFDs (r^2=.184). Fisher's r-to-z test revealed a significant difference in explained variance (z=1.95; P = 0.05).*

1.6. Discussion and Conclusion

We have examined the utility of three 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, topic models and recurrent neural network models would account for the predictability of a token, given the preceding tokens in the sentence, as perceived by humans, as well as the some electroencephalographic and eye movement data that are typically explained by it. Therefore, we used the amount of explained item-level variance as a benchmark, which has been established as standard evaluation criterion for neurocognitive models of visual word recognition (e.g., [HOF 08; PER 10; SPIE 97]).

Our hypothesis was at least partially confirmed: n-gram models and RNNs, 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. This, however, might at least in part be explainable

by a larger amount of noise in the EEG data with less participants as compared to the eye movement data with much more participants.

The long-range semantic relationships as captured by topic models on the other hand, provided a different picture. If any, the topic model made only a minor contribution to predictability and the N400. For the fast and successful recognition of a word at one glance, as reflected by SFDs in contrast, long-range document level relationships seem to provide a stronger contribution. This result pattern occurs even in the context of single sentences, without a discourse level setting the topic of a document. This suggests that colloquial and taxonomic far-reaching semantic long-term structure particularly determines the fast and effective single-glance recognition of a word within the first 300 ms after the onset of word recognition, but hardly somewhat later processes around 400 ms in the brain-electric data and the time-consuming probably late processes being contained in the predictability scores.

For predicting the empirical word predictability from sentence context as well as the N400, recurrent neural network models performed often somewhat worse than the n-gram approach. For predicting SFDs, however, the neural model was superior. Most interestingly, the neural network model performs best when it is trained on a small but probably more representative sample of everyday language. So size does probably not trump everything and in any model [BAN 01]. It also hints at the generalization properties of its dimensionality reduction, which are more important for smaller training data [BUL 07; GAM 16; HOF 14], but probably leads to imprecise modeling when more data is available.

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 Figure 1.2; cf. Figures in Biemann et al., 2015).

When contrasting the standard predictors of position, frequency and predictability used in eye tracking and EEG research (e.g., [DAM 06; REI 03]), only for the SFDs all three corpus-based predictors did a better job than the standard model. However, with this approach it is apparent that many more predictors are needed, and thus the probability for fitting error variance is much larger than for the standard model. Thus we think that much more evidence is required, before we dare to state this as a firm conclusion. Also for this three-predictor model, adding the empirical predictability is providing a net gain of 2% explained variance.

Because n-gram or neural 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, topic models and neural models can be used to generalize computational models of eye movement control to novel sentences [ENG 05; REI 03].

In the end, language models can also improve our understanding of the cognitive processes underlying predictability, EEG, and EM measures. While it is not clear what exactly determines human CCP-based predictability performance, the different language models provide differential grain size levels applied their training data, thus paving the the way for the question which neurocognitive measures of 'word predictability' are affected by sentence- or document-level semantic knowledge. While Ziegler and Goswami [ZIE 05] discussed the optimal grain size of language learning at the word-level and sub-word-level grain sizes, recent evidence of a severe decline of comprehension abilities since the 60s suggests the necessity to continue with that discussion at the level of supralexical semantic integration [SPI 16].

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

[BAA 95] BAAYEN H. R., PIEPENBROCK R., GULIKERS L., *The CELEX Lexical Database. Release 2 (CD-ROM)*. LDC, University of Pennsylvania, Philadelphia.

[BAA 10] BAAYEN H.R. "Demythologizing the word frequency effect: A discriminative learning perspective", *The Mental Lexicon*, vol. 5 no. 3, 2010, p. 436-461.

[BAN 01] BANKO M., BRILL E. "Scaling to very very large corpora for natural language disambiguation", *Proc. ACL '01*, Toulouse, France, 2001, p. 26-33.

[BAR 07] BARBER H. A., KUTAS M. "Interplay between computational models and cognitive electrophysiology in visual word recognition", *Brain Res. Rev.*, vol. 53 no. 1, 2007, p. 98-123.

[BEN 03] BENGIO Y., DUCHARME R., VINCENT P., JAUVIN C. "A neural probabilistic language model", *Journal of Machine Learning Research*, vol. 3 no. 6, 2003.

[BIE 12] BIEMANN C., ROOS S., WEIHE K. "Quantifying semantics using complex network analysis", *Proc. COLING 2012*, Mumbai, India, 2012, p. 263-278.

[BIE 15] BIEMANN C., REMUS S., HOFMANN M. J. "Predicting word 'predictability' in cloze completion, electroencephalographic and eye movement data", *Proceedings of the 12th International Workshop on Natural Language Processing and Cognitive Science*. Krakow, Poland, 2015.

[BLE 03] BLEI D. M., NG A. Y., JORDAN M. I. "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, 2003, p. 993-1022.

[BLO 80] BLOOM P. A., FISCHLER I. "Completion norms for 329 sentence contexts", *Memory & cognition*, vol. 8 no. 6, 1980, p. 631-642.

[BUL 07] BULLINARIA J. A., LEVY J. P. "Extracting semantic representations from word co-occurrence statistics: a computational study", *Behavior research methods*, vol. 39, no. 3, 2007, p. 510–26.

[BRY 11] BRYSBAERT M., BUCHMEIER M., CONRAD M., JACOBS A. M., BÖLTE J., BÖHL A. "A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German", *Experimental psychology*, vol. 58, 2011, p. 412-424.

[COO 10] COOLICAN H. "Research Methods and Statistics in Psychology", *Hodder & Stoughton*, 2010

[DAM 07] DAMBACHER M., KLIEGL R. "Synchronizing Timelines: Relations between fixation durations and N400 amplitudes during sentence reading", *Brain research*, vol. 1155, 2007, p. 147-162.

[DAM 06] DAMBACHER M., KLIEGL R., HOFMANN M. J., JACOBS A. M. "Frequency and predictability effects on event-related potentials during reading". *Brain research*, vol. 1084 no. 1, 2006, p. 89-103.

[DAM 09] DAMBACHER A. M. "Bottom-up and top-down processes in reading", *Universitätsverlag Potsdam*, Potsdam, 2009.

[ELM 90] ELMAN J. L., "Finding Structure in Time," *Cognitive Science*, vol. 211, 1990, p. 1-28.

[ENG 05] ENGBERT R., NUTHMANN A., RICHTER E. M., KLIEGL R. "SWIFT: a dynamical model of saccade generation during reading", *Psychological review*, vol. 112 no. 4, 2005, p. 777-813.

[GAM 16*] GAMALLO P. "Comparing explicit and predictive distributional semantic models endowed with syntactic contexts," *Language Resources and Evaluation*, *to appear 2016

[GOL 12] GOLDHAHN D., ECKART T., QUASTHOFF U. "Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages", *Proc. LREC 2012*, Istanbul, Turkey, 2012, p. 759-765

[GRI 07] GRIFFITHS T. L., STEYVERS M., TENENBAUM J. B. "Topics in Semantic Representation", *Psychological review*, vol. 114 no. 2, 2007, p. 211-244.

[HOF 14] HOFMANN M. J., JACOBS A. M. "Interactive activation and competition models and semantic context: From behavioral to brain data". *Neuroscience and biobehav. rev.s*, vol 46, 2014, p. 85-104.

[HOF 08] HOFMANN M. J., TAMM S., BRAUN M. M., DAMBACHER M., HAHNE A., JACOBS A. M. "Conflict monitoring engages the mediofrontal cortex during nonword processing", *Neuroreport*, vol. 19 no. 1, 2008, p. 25-29.

[KLI 04] KLIEGL R., GRABNER E., ROLFS M., ENGBERT R. "Length, frequency, and predictability effects of words on eye movements in reading", *Europ. Journal of Cog. Psy.*, vol. 16 no. 12, 2004, p. 262-284.

[KNE 95] KNESER R., NEY H. "Improved backing-off for m-gram language modeling", *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, Detroit, Michigan, 1995, p. 181-184.

[KUT 11] KUTAS M., FEDERMEIER K. D. "Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)", *Ann. Rev. of Psychology*, vol. 62, 2011, p. 621-647.

[KUT 84] KUTAS M., HILLYARD S. A. "Brain potentials during reading reflect word expectancy and semantic association", *Nature*, vol. 307 no. 5947, 1984, p. 161-3.

[LAN 97] LANDAUER T. K., DUMAIS S. T. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", *Psychological Review*, vol. 104 no. 2, 1997, p. 211-240.

[MAN 99] MANNING C. D., SCHÜTZE H. *Foundations of Statistical Natural Language Processing*, Cambridge, MA, USA, MIT Press, 1999.

[MAR 71] MARR D. "SIMPLE MEMORY: A THEORY", *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 262 no. 841, 1971, p. 23-81.

[MCC 81] MCCLELLAND J. L., RUMELHART D. E. "An Interactive Activation Model of Context Effects in Letter Perception: Part 1", *Psychological Review*, vol. 5, 1981, p. 375-407.

[MIK 12] MIKOLOV T. "Statistical language models based on neural networks", PhD thesis, Brno University of Technology, 2012.

[MIK 13] MIKOLOV T., YIH W., ZWEIG G. "Linguistic Regularities in Continuous Space Word Representations", *Proceedings of NAACL-HLT*, Atlanta, GA, USA, 2013, p. 746-751.

[NOR 03] NORMAN K. A., O'REILLY R. C. "Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach.," *Psychological Review*, vol. 110 no. 4, 2003, p. 611-46.

[PER 10] PERRY C., ZIEGLER J. C., ZORZI M. "Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model", *Cognitive Psychology*, vol. 61 no. 2, 2010, p. 106-51.

[PHA 07] PHAN X-H., NGUYEN C-T. *GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA)*, http://gibbslda.sourceforge.net/, 2007

[RAD 12] RADACH R., GÜNTHER T., HUESTEGGE L. "Blickbewegungen beim Lesen, Leseentwicklung und Legasthenie", *Lernen und Lernstoerungen*, vol. 1 no. 3, 2012, p. 185-204.

[REI 03] REICHLE E. D., RAYNER K., POLLATSEK A. "The E-Z reader model of eye-movement control in reading: comparisons to other models", *The Behavioral and brain sciences*, vol. 26 no. 4, 2003, p. 445–476; discussion p. 477-526.

[SPI 16*] SPICHTIG A., HIEBERT H., VORSTIUS C., PASCOE J., PEARSON P., RADACH R. "The Decline of Comprehension-Based Silent Reading Efficiency in the U.S.: A Comparison of Current Data with Performance in 1960", *Reading Research Quarterly*, to appear 2016.

[SPI 97] SPIELER D. H., BALOTA D. A. "Bringing Computational Models of Word Naming Down to the Item Level". *Psychological Science*, vol. 8 no. 6, 1997, p. 411-416.

[TAY 53] TAYLOR W. L. "'Cloze' procedure: A new tool for measuring readability", *Journalism Quarterly*, vol. 30, 1953, p. 415.

[ZIE 05] ZIEGLER J. C., GOSWAMI U. "Reading Acquisition , Developmental Dyslexia , and Skilled Reading Across Languages : A Psycholinguistic Grain Size Theory", *Psychological Bulletin*, vol. 131 no. 1, 2005, p. 3-29.