Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing

Christoph Aurnhammer (aurnhammer@coli.uni-saarland.de)
Department of Language Science and Technology, Saarland University, Campus C1, 66123 Saarbrücken, Germany
Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands

Abstract

The Simple Recurrent Network (SRN) has a long tradition in cognitive models of language processing. More recently, gated recurrent networks have been proposed that often outperform the SRN on natural language processing tasks. Here, we investigate whether two types of gated networks perform better as cognitive models of sentence reading than SRNs, beyond their advantage as language models. This will reveal whether the filtering mechanism implemented in gated networks corresponds to an aspect of human sentence processing. We train a series of language models differing only in the cell types of their recurrent layers. We then compute word surprisal values for stimuli used in self-paced reading, eye-tracking, and electroencephalography experiments, and quantify the surprisal values' fit to experimental measures that indicate human sentence reading effort. While the gated networks provide better language models, they do not outperform their SRN counterpart as cognitive models when language model quality is equal across network types. Our results suggest that the different architectures are equally valid as models of human sentence processing.

Keywords: Surprisal; Gated Recurrent Neural Networks; Language Modeling; Sentence Processing; Sentence Reading; Self-paced Reading; Eye-tracking; Electroencephalography

Introduction

In psycholinguistics, the Simple Recurrent Network (SRN; Elman, 1990) has been a popular (and reasonably successful) neural architecture for modeling aspects of human sentence processing, and it remains so to this day (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Frank, Otten, Galli, & Vigliocco, 2015; Rabovsky, Hansen, & McClelland, 2018; Twomey, Chang, & Ambridge, 2014, to name just a few recent examples). However, it has been known since the late 1990s that the SRN struggles to integrate information over many classification steps, due to what is referred to as the vanishing gradient problem (Hochreiter, 1998).

This problem was addressed by neural network models containing recurrent units that have gates with trained weights, such as the Gated Recurrent Unit (GRU; Bahdanau, Cho, & Bengio, 2015) and the Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) network. The gating mechanism implemented in GRUs and LSTMs controls the flow of information in the recurrent cell, allowing the cells to memorise information over time, forget it when adequate, and to determine the weighting of old and new input. While the principles of the two architectures are similar, the GRU can be regarded as a more lightweight variation on the LSTM, making use of only two gates and a single hidden state, whereas the LSTM architecture provides three gates and introduces an additional memory state.

Gated networks outperform SRNs on several NLP tasks. For example, LSTMs perform more accurately than SRNs on number agreement (Linzen, Dupoux, & Goldberg, 2016) and conversational speech recognition (Xiong et al., 2017). In the current study, we investigate how well gated networks perform as cognitive models of human sentence processing compared to the traditional SRN. We model human word-level processing effort by using recurrent neural networks as probabilistic language models that estimate the predictability of words in context.

For the language modeling problem, the ability to make effective use of more of the words in the prior sequence can be expected to pose a crucial advantage of a gated recurrent network compared to the SRN. For instance, the processing of long-term dependencies has been proposed as one aspect of natural language processing addressed more adequately by gated networks than by SRNs (Bahdanau et al., 2015). Because gated networks are designed for long-distance encoding, they may also be superior cognitive models: The filtering mechanism implemented by the gates may mirror an aspect of human sentence processing. For example, it is known that humans read the word *or* faster when they processed the word *either* in the prior sequence of words, demonstrating their ability to remember dependencies between words across long spans (Staub & Clifton Jr., 2006). Gated networks may reflect this human behaviour more accurately than SRNs by assigning lower surprisal to the word *or* even when the corresponding *either* is distant.

Although LSTMs and GRUs have already been applied to account for human language performance measures (Futrell et al., 2019; Goodkind & Bicknell, 2018; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Hahn & Keller, 2016; McCoy, Frank, & Linzen, 2018; Sakaguchi, Duh, Post, & Durme, 2017; Van Schijndel & Linzen, 2018a, 2018b), the question remains whether they form more accurate cognitive processing models than traditional SRNs, beyond what might be expected from their stronger language modeling abilities.

In the current study, we directly compare three recurrent neural network (RNN) language model architectures (SRNs, GRUs, and LSTMs) on their ability to predict human reading data collected in self-paced reading, eye-tracking, and electroencephalography experiments. If the mechanisms implemented in GRUs and LSTMs correspond to cognitive mechanisms applied during sentence comprehension, we would expect predictions by these models to fit human reading data more closely than predictions by SRNs, over and above any advantage that GRUs/LSTMs might have because of their superiority as language models. Conversely, if the cognitive system does not apply anything like a gating mechanism, the SRN may simulate human language processing more closely than GRUs and LSTMs do. In that case, the SRN may predict human processing data more accurately than gated RNNs that are matched for language model quality.

Method[1]

To determine whether or not LSTMs and GRUs outperform SRNs as cognitive models of sentence processing, we train three different kinds of RNN language models, each using one of the three recurrent cell types. We evaluate the models by assessing the predictive power of the surprisal values they assign to stimuli used in three experiments of humans sentence reading.

Human processing data

We assess how well each RNN language model's word surprisal values predict human cognitive processing effort during sentence reading, as measured in self-paced reading (SPR), eye-tracking (ET), and electroencephalography (EEG) experiments. The SPR and ET data come from Frank, Monsalve, Thompson, and Vigliocco (2013) and the EEG data from Frank et al. (2015).

In all three experiments, participants read English sentences sampled from unpublished novels. All sentences are understandable out of their context in the novels. A subset of the sentences were used in the ET and EEG experiments; these were the shortest sentences (maximum length: 15 words) of those from the SPR study. Table 1 displays the numbers of participants and stimuli, along with ranges and means of sentence length for each of the three data sets. Importantly, we make sure that all word types in the stimuli are attested for in the training data, meaning that the language models do not encounter words for the first time when applied to the stimuli.

For this study, we select a single variable from each dataset that is indicative of human processing cost: Reading time (RT) from the SPR data, gaze duration (a.k.a. first-pass reading time) from the ET data, and N400 size from the EEG data set. We follow the insight that reading times reflect the cognitive effort the reader needs to employ during language processing (Levy, 2008). Reflecting this idea, the N400 event-related potential amplitude indicates processing effort on lexico-semantic levels (Kutas, Van Petten, & Kluender, 2006; Kutas & Federmeier, 2011). Earlier research has already demonstrated that these dependent variables, from these particular data sets, indeed correlate with word surprisal values (SPR: Monsalve, Frank, & Vigliocco, 2012; ET: Frank & Thompson, 2012; EEG: Frank et al., 2015).

Network architectures

Our RNN architecture consists of a 400-unit word embedding layer, a 500-unit recurrent layer, a 400-unit feed-forward layer with tanh activation function, and a final layer with log-softmax activation function, which maps to the vocabulary. We do not use pre-trained word embeddings. Rather, the weights of the embedding layer that transforms the vocabulary items to real-valued word vectors are learned during the next-word prediction task, along with the rest of the network weights. The model architectures only differ in that their recurrent layers use either SRN, LSTM, or GRU cells.

Training corpus

As training data for the language models we use section 13 of the English version of the Corpora from the Web (COW, 2014 version; Schäfer, 2015). This corpus consists of randomly ordered sentences collected from web pages. From this section, the 10,000 most frequent word types are selected as our model's vocabulary. One hundred and three word types that appear in the experimental stimuli (see Section on *Human Processing Data*) but are not yet covered in the vocabulary are added, resulting in a final vocabulary size of 10,103 word types. After determining the vocabulary, we select those sentences from the initial COW section that contain only in-vocabulary word types, thus also covering the low-frequency words in the experimental stimuli. We follow this strategy to avoid having to use a (cognitively implausible) UNKNOWN-type. Furthermore, we only keep sentences with a maximum length of 39 words, which corresponds to the longest sentence in the experimental stimuli (not counting punctuation as words). We remove a small number of sentences to arrive at a final selection that contains 6,470,000 training sentences and consists of 94,422,754 tokens in total.

Although this training set and vocabulary size is relatively small by current standards, note that our aim here is not to construct the best possible language model, and not even to provide the most accurate account of human sentence processing effort. Rather, we investigate whether RNN architectures differ in their ability to predict human data.

Network training

We train the networks on one sentence at a time to let model training resemble human language processing and acquisition. Further, we reset the hidden state of the recurrent cells to zero for each new sentence. From the network's log-probability output at each step, the loss function computes the negative log-likelihood. Based on this loss, we optimise the network weights using stochastic gradient descent with momentum (0.9) and an initial learning rate of 0.0025. After each third of the training data, we reduce the learning rate to half of its prior value. As precaution to the exploding gradient problem (Bengio, Simard, & Frasconi, 1994), we clip gradients at 0.25. The error is always back-propagated through the

[1] All code and data is available at `https://github.com/caurnhammer/AurnhammerFrank_CogSci2019`

Table 1: Numbers of participants, number of sentences, range of sentence length, mean sentence length, number of word tokens, and number of data points (after exclusion; see Section *Stage 1: Predicting human data from surprisal*) in the human sentence reading data sets. In the SPR experiment, each participant received a random subset of the 361 possible sentences (see Frank et al., 2013, for details).

Exp.	Part.	Sent.	Range sent. len.	Mean sent. len.	Tokens	Data points
SPR	54	361	5–39	14.1	4957	132,858
ET	35	205	5–15	9.4	1931	28,970
EEG	24	205	5–15	9.4	1931	24,618

entire sentence.

To account for random variation in model performance that is solely due the initial weights and training sentence presentation order, we train each RNN type six times, each time with different random initial weights (uniformly distributed between ± 0.1; with initial biases 0) and a different random order of sentence presentation. However, for each training repetition, the same initial weights (for connections that correspond between architectures) and the same presentation orders are applied across the three recurrent cell architectures. Hence, the *only* difference between the RNN types is in the architectures of their recurrent cells.

Language model evaluation

We evaluate the performance at the nine different training corpus sizes by computing the perplexity on the unseen experimental stimulus sentences. Perplexity is computed as

$$PPL = e^{-|W|^{-1} \sum_{w \in W} \log P(w)},$$

where $|W|$ is the number of word tokens in the experimental sentences. Lower perplexity results from language models that assign higher probabilities to the test data. Perplexity thus expresses the extent to which a language model captures the statistical structures of the data that are useful to predicting the next word, irrespective of the extent to which this is helpful for explaining human sentence processing measures.

Statistical model evaluation

The RNN models' ability to account for the human processing data is evaluated in two stages, as explained in more detail below. First, we compute surprisal for the experimental test items. Surprisal is computed as

$$\text{surprisal}(w_t) = -\log\ P(w_t | w_1, ..., w_{t-1}).$$

and formalises the extent to which occurrence of a word w_t is unexpected, given a sequence of preceding words $w_1, ..., w_{t-1}$ (Hale, 2001; Levy, 2008). The reading-time and N400 measures on each word are regressed on each model's surprisal estimates resulting in a collection of goodness-of-fit measures. Next, we assess the relation between each RNN type's goodness-of-fit and its quality as a language model.

Stage 1: Predicting human data from surprisal Each individual RNN generates surprisal estimates for each word of the 361 stimuli sentences. The surprisal values are obtained after training the network on 1K, 3K, 10K, 30K, 100K, 300K, 1M, 3M, and all 6.47M sentences. This procedure allows to observe how the goodness-of-fit to human data develops as a function of language model quality, which steadily increases with the amount of observed training data. In summary, we have 9 (points during training) × 6 (training repetitions) × 3 (RNN types) = 162 sets of surprisal values to compare to the SPR times, gaze durations, and N400 sizes.

The predictive power of each set of surprisal values is assessed by means of linear mixed effects regression, using the `MixedModels` package[2] (v0.18.1) for `Julia` (Bezanson, Edelman, Karpinski, & Shah, 2017). First, a baseline model was fitted to each of the three human data sets. The aim of this baseline is to factor out the effects of the most imporant variables known to affect reading times and N400 sizes and thus be left with an effect of surprisal that is as isolated as possible.

The dependent variables self-paced reading times and gaze durations are log-transformed. In the EEG data, N400 size is analysed as defined by Frank et al. (2015): the average potential on central-parietal electrodes over a 300–500ms window after word onset.

The baseline models include as fixed effects: log-transformed word frequency in the training corpus, word length (number of characters) and word position in the sentence. For the SPR and ET data, we also enter the previous word's frequency and length into the analysis to account for spillover effects that are known to affect reading times (Rayner, 1998). Moreover, we add previous-word RT (log-transformed) to the SPR analysis to address the high correlation between consecutive word RTs that typically occurs in the SPR paradigm; and to the ET analysis we add a binary factor indicating whether the previous word was fixated. For the EEG analysis, we enter baseline activity (i.e., the average electrode potential in the 100ms leading up to word onset) into the regression. All interactions between the fixed effects are also included. Furthermore, there are by-subject and by-item (word token) random intercepts and by-subject random slopes of all fixed-effect predictors.

We exclude data on sentence-initial and -final words, words attached to a comma, and clitics. Furthermore, participants are removed from the analysis if they are not native English

[2] `github.com/dmbates/MixedModels.jl`

speakers or scored less than 80% correct on the yes/no comprehension questions that were presented for approximately half the sentence stimuli. In addition, SPR and ET data points are removed on words directly following a comma or clitic, and when reading times are below 50ms or over 3500ms. For the EEG data, we exclude artefacts as identified by Frank et al. (2015).

The goodness-of-fit of each set of surprisal values for each human data set equals the log-likelihood ratio (decrease in regression model deviance) between the baseline and a regression model that additionally includes surprisal as both a fixed effect and by-subject random slope. For the SPR and ET analyses, the previous word's surprisal is also added (again as fixed and random effects) in order to capture spillover effects. The resulting values are χ^2-statistics, with 2 degrees of freedom for the EEG data and 4 degrees of freedom for the two reading-time data sets. We further add a negative sign to the χ^2-statistics to indicate effects in the negative-going direction, that is, when higher surprisal results in shorter reading times or smaller (less negative) N400 size.

Stage 2: Predicting goodness-of-fit from language model accuracy Networks that form better language models tend to estimate surprisal values that fit human data better (Frank et al., 2015; Goodkind & Bicknell, 2018). In analysis Stage 2, we are interested in ascertaining whether the relation between language model accuracy and goodness-of-fit to human data differs between network architectures.

We quantify language model accuracy as the average log-probability (i.e., negative average surprisal) estimated over the experimental sentences, weighted by the number of times each word token takes part in the analysis described above, that is, for how many participants the data on this word was not excluded. Following this, we fit Generalized Additive Mixed Models (GAMMs), for each of the three RNN types and human data sets separately, to predict the goodness-of-fit measures (from analysis Stage 1) from the language model accuracies, with network training repetition as a random effect. This is done using the `R` package `mgcv` (Wood, 2004).

Results

Language modelling results

Figure 1 reports on the perplexities of the 18 individual language models at 9 different points during training. While the SRNs set in at lower perplexity than the gated networks early in training, the latter ultimately outperform the simple RNNs. Language model performance steadily increases throughout training but a saturation of the language model performance seems only to commence at the final training steps.

Statistical modelling results

Figure 2 displays the goodness-of-fit measures from analysis Stage 1 for each human data set, as well as the fitted curves relating goodness-of-fit to language model accuracy from analysis Stage 2. These plots clearly show that well-trained language models estimate surprisal values that account for reading times and N400 size, and that the goodness-of-fit generally improves as the language models more accurately capture the linguistic patterns. Interestingly, for lower levels of linguistic accuracy, corresponding to models trained on relatively few sentences, the effect of surprisal on gaze duration size is reversed, in that higher surprisal correlates with faster reading. The cause of this reversal remains to be identified.

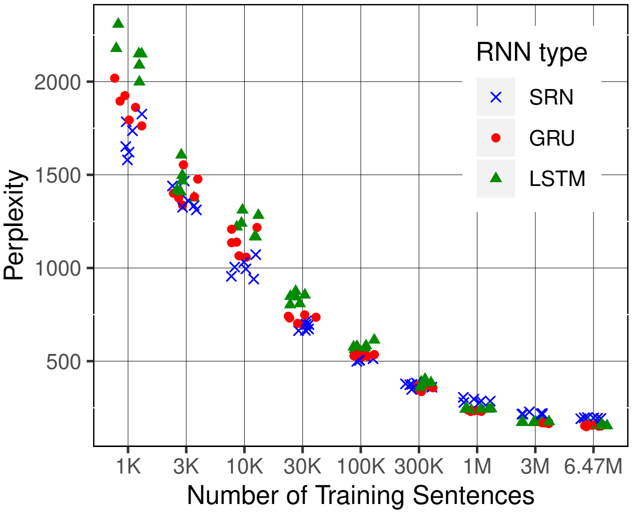

Figure 1: Perplexity on the experimental sentences for each of three RNN types at nine different training corpus sizes. At each training size there are six models of each type with different sentence orderings and initial weights. Data points are subjected to horizontal jitter to improve readability.

The gated RNN models reach higher levels of language model accuracy than the SRNs, which is why they can also outperform SRNs in terms of goodness-of-fit. For similar levels of language model accuracy, however, the three model types account for similar quantities of variance in the human processing data, as is evident from the largely overlapping confidence intervals of the fitted GAMM curves.

This does not imply that different network types make no independent contributions to human data prediction. To test whether the models differ qualitatively in that one RNN explains unique variance over and above the others, we average the surprisal values over the six fully trained versions of each network architecture. Next, we fit linear mixed models including the surprisals from two of the three RNN types and then test whether that regression model fits the data better than a regression with only a single set of surprisal values. That is, for each pair of RNN types we ask whether one explains human data over and above the other.

Table 2 shows model comparisons, testing for the significance of adding the surprisal from the models displayed in rows to the models in columns. The comparisons reveal statistically significant effects of GRU and LSTM surprisal over and above SRN surprisal in all three data sets. For the EEG data, SRN surprisal also explains variance not yet explained

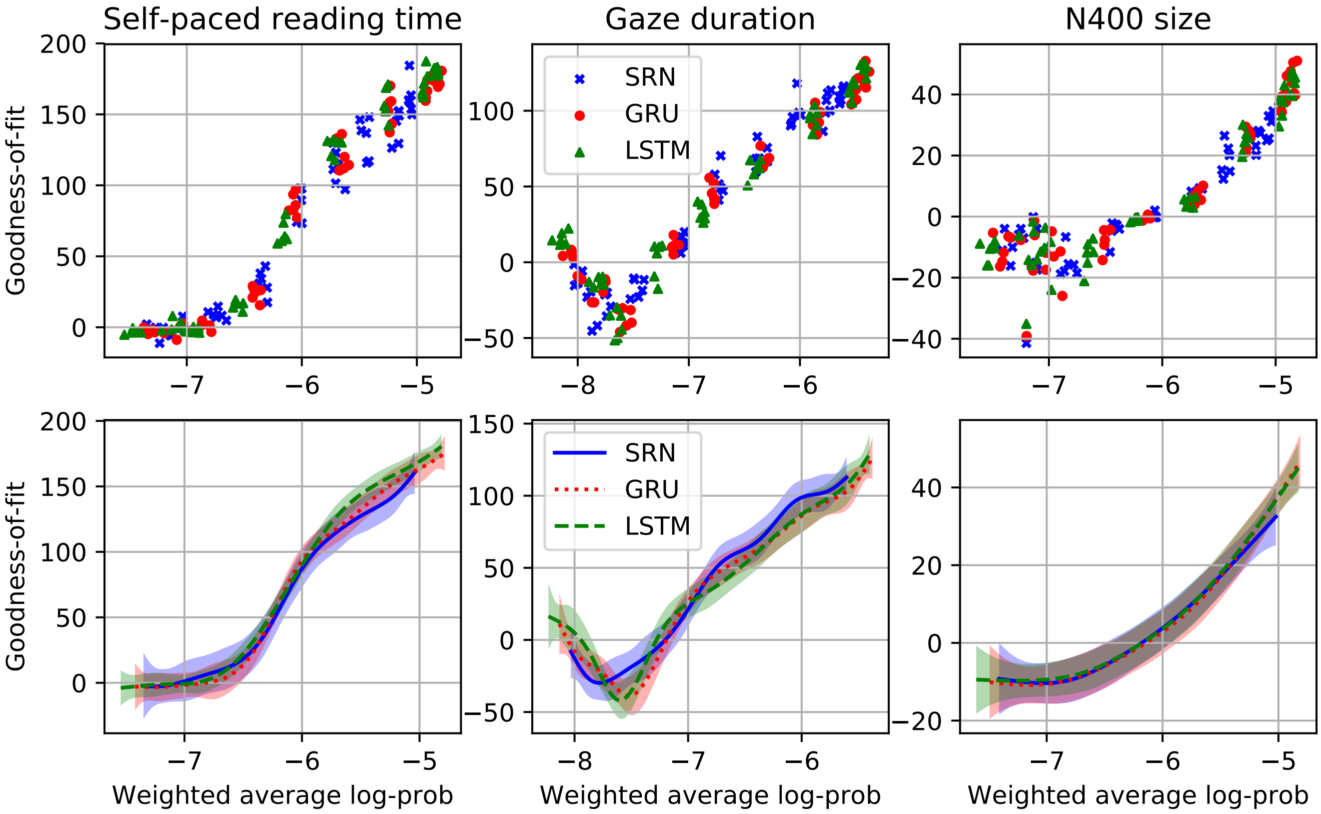

Figure 2: Top row: results from analysis Stage 1. The goodness-of-fit of surprisal to human data is plotted as a function of language model accuracy. Bottom row: results from analysis Stage 2. Plotted are the fitted GAMM curves relating goodness-of-fit to language model accuracy. Shaded areas indicate 95% confidence intervals. Panels on the left, middle, and right side are for SPR, ET, and EEG data, respectively.

by the gated networks.

Discussion

Our comparison of the abilities of SRNs, GRUs, and LSTMs to predict human reading-time and N400 measures (via the networks' word-surprisal estimates) do not reveal any large or reliable difference between the three RNN types, at least, not as long as the different networks' accuracies as language models do not differ. The two gated networks do form better language models than the SRN, resulting in more precise predictions of human data at the highest levels of language model accuracy. However, if the human cognitive system would employ mechanisms akin to the gates in GRU/LSTM recurrent cells, we would expect GRU/LSTM-based surprisal to show better fit than SRN-based surprisal to the human processing data, even without any difference in language model accuracy. Our analyses do not support this conclusion.

The gated RNNs explain variance over and above what is accounted for by the SRNs on SPR, ET, and EEG data. This is an expected effect, given that gated networks form better language models. Their ability to encode relations between word tokens along larger spans is likely giving them a clear advantage in accounting for human data. More surprisingly, on the EEG data the SRNs also explain a portion of variance that is distinct from the one explained by the gated networks. This finding may suggest a potential insensitivity of the N400 ERP component to long-distance dependencies, at least to the extent that N400 size reflects word predictability. Converging evidence for this interpretation is presented by Frank et al. (2015) who demonstrate that an *n*-gram language model with a context size of three words explains variance over and above an SRN on the same data set.

Conclusion

While gated recurrent neural networks provide better language models than simple recurrent networks, our investigations do not indicate that they have any substantial or reliable advantage as cognitive models of sentence reading, in addition to what is expected from their superior language modeling abilities. Nevertheless, gated networks consistently reached higher linguistic accuracy. This fact alone makes the use of gated RNN advisable not only from a language modeling point of view but also for psycholinguistics (and cognitive science more in general) when as much variance in human data as possible needs to be explained, for example when surprisal is used as a covariate in studies that aim to find a unique

Table 2: Results from regression model comparisons between RNN types. Each χ^2-statistic is the outcome of a log-likelihood ratio test for whether the network type in the table row accounts for variance in the human data over and above the network type in the table column. Asterisks indicate statistical significance level after multiple-comparison correction (Benjamini & Hochberg, 1995): * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Exp.		SRN	GRU	LSTM
SPR	SRN		$\chi^2(4) = 3.20$	$\chi^2(4) = 3.69$
	GRU	$\chi^2(4) = 12.7$*		$\chi^2(4) = 1.29$
	LSTM	$\chi^2(4) = 18.1$**	$\chi^2(4) = 6.22$	
ET	SRN		$\chi^2(4) = 6.18$	$\chi^2(4) = 8.70$
	GRU	$\chi^2(4) = 15.6$*		$\chi^2(4) = 0.46$
	LSTM	$\chi^2(4) = 22.5$***	$\chi^2(4) = 4.88$	
EEG	SRN		$\chi^2(2) = 10.9$*	$\chi^2(2) = 8.65$*
	GRU	$\chi^2(2) = 26.0$***		$\chi^2(2) = 3.26$
	LSTM	$\chi^2(2) = 21.7$***	$\chi^2(2) = 1.22$	

effects of some additional predictor.

Acknowledgments

The work presented here was funded by SFB 1102 "Information density and linguistic encoding", awarded by the German Research Foundation (DFG); and by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations.* San Diego, CA: ICLR.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*(S6), 1318-1352.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frank, S. L., & Thompson, R. L. (2012). Early effects of word surprisal on pupil size during reading. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1554–1559). Austin, TX: Cognitive Science Society.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Salt Lake City, UT: CMCL.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* New Orleans, Louisiana: Association for Computational Linguistics.

Hahn, M., & Keller, F. (2016). Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 85–95). Austin TX: Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In *Handbook of psycholinguistics (second edition)* (pp. 659–724). Elsevier.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceeding of the 40th Annual Conference of the Cognitive Science Society* (pp. 2093–2098). Madison, WI: Cognitive Science Society.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408).

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*, 693–705.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Sakaguchi, K., Duh, K., Post, M., & Durme, B. V. (2017). Robsut wrod reocginiton via semi-character recurrent neural network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 3281–3287). San Francisco, CA: AAAI.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora* (pp. 28–34). Mannheim, Germany: Institut für Deutsche Sprache.

Staub, A., & Clifton Jr., C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425–436.

Twomey, K. E., Chang, F., & Ambridge, B. (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology*, *73*, 41–71.

Van Schijndel, M., & Linzen, T. (2018a). Modeling garden path effects without explicit hierarchical syntax. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2603–2608). Austin, TX: Cognitive Science Society.

Van Schijndel, M., & Linzen, T. (2018b). A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4704–4710). Brussels: Association for Computational Linguistics.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., … Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2410–2423.