Evaluating information-theoretic measures of word prediction in naturalistic sentence reading

Aurnhammer Christoph^a, Stefan L. Frank^{b,*}

^a Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

^b Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:

Predictive processing
Sentence processing
Next-word entropy
Surprisal
Lookahead information gain
Reading time
Electroencephalography
Psycholinguistics
Computational linguistics

ABSTRACT

We review information-theoretic measures of cognitive load during sentence processing that have been used to quantify word prediction effort. Two such measures, *surprisal* and *next-word entropy*, suffer from shortcomings when employed for a predictive processing view. We propose a novel metric, *lookahead information gain*, that can overcome these short-comings. We estimate the different measures using probabilistic language models. Subsequently, we put them to the test by analysing how well the estimated measures predict human processing effort in three data sets of naturalistic sentence reading. Our results replicate the well known effect of surprisal on word reading effort, but do not indicate a role of next-word entropy or lookahead information gain. Our computational results suggest that, in a predictive processing system, the costs of predicting may outweigh the gains. This idea poses a potential limit to the value of a predictive mechanism for the processing of language. The result illustrates the unresolved problem of finding estimations of word-by-word prediction that, first, are truly independent of perceptual processing of the to-be-predicted words, second, are statistically reliable predictors of experimental data, and third, can be derived from more general assumptions about the cognitive processes involved.

1. Introduction

Over the last 15 years,¹ the role of prediction during human language processing has fueled many debates in cognitive science and linguistics. In psycholinguistics, the idea of prediction has become a key explanation to the question of how humans process language so efficiently, for instance at “rates of 120–200 words per minute” (Broderick et al., 2017, p. 803) during speech comprehension. A general principle of prediction in language comprehension has strong explanatory power for many phenomena that have previously been observed in experimental studies, such as reduced reading time (Staub, 2015) and N400 strength (Kutas and Federmeier, 2011) for words that can be expected from prior sentence context, as well as for efficient turn taking in dialogue (Corps et al., 2018), to name only a few examples.

Theoretical accounts engaging with the anticipation of upcoming linguistic elements differ in the importance they attribute to predictive mechanisms for language processing, with some regarding the brain essentially as a proactive prediction machine (Bar, 2007; Clark, 2013;

Friston, 2010; den Ouden et al., 2012), while others argue that prediction occurs under only under specific circumstances (Huettig, 2015) or even question the usefulness of a predictive mechanism for the processing of language (e.g., Jackendoff, 2007).

In the present work, we are particularly interested in the phenomenon of word prediction (as opposed to, e.g., prediction of syntactic structure; Staub and Clifton, 2006), a phenomenon supported by a large body of work (Dambacher et al., 2009; DeLong et al., 2005; Dikker et al., 2010; Dikker and Pykkänen, 2013; Federmeier, 2007; Laszlo and Federmeier, 2009; Lau et al., 2014; van Berkum et al., 2005; Wicha et al., 2004).² In particular, we investigate whether the upcoming word is probabilistically predicted (anticipated) from prior words, even before perceptual processing of the upcoming word commences.

Methodologically, this study follows a line of work that employs probabilistic language models (PLMs) to quantify cognitive load during human language processing. Trained on large text collections, PLMs assign probabilities to next words when presented with a sequence of prior words. Notably, the predictions these models make correlate with

* Corresponding author. P.O. Box 9103, 6500 HD, Nijmegen, the Netherlands.

E-mail addresses: aurnhammer@coli.uni-saarland.de (C. Aurnhammer), s.frank@let.ru.nl (S.L. Frank).

¹ Although the idea is much older, see Huettig (2015).

² A large-scale replication attempt (Nieuwland et al., 2018) demonstrated limitations of the results of DeLong et al. (2005), suggesting there is at most a small effect. See also the replication in Nicenboim et al. (2019) using even more data and the modeling approach in Rabovsky (2019).

reading times (Goodkind and Bicknell, 2018; Hahn and Keller, 2016; Monsalve et al., 2012; Smith and Levy, 2013), N400 sizes (Frank et al., 2015), and voxel activation in the brain (Hale et al., 2015; Willems et al., 2016) during sentence or text comprehension.

For the evaluation of PLM-based measures of cognitive effort, it is common to collect human processing data over all words of naturally occurring sentences, rather than looking only at critical words of carefully crafted experimental items. Probabilities are then estimated over the same words and potential confounds are factored out in a large-scale regression analysis. This is also the approach we take here.

We study three sets of human data collected using EEG, eye-tracking, and self-paced reading while participants read naturalistic sentences. In the EEG data, we focus on the N400 event-related potential amplitude, an electrophysiological measurement of word level processing that is often interpreted as an index of predictive processing (van Petten and Luka, 2012). Similarly, both self-paced reading times and first-pass durations (measured in eye-tracking) have been found sensitive to predictive processes (van Berkum et al., 2005; Staub and Clifton, 2006).

In Section 2, we first describe information-theoretic metrics that have been proposed as measures of comprehension effort and discuss their interpretation in the context of predictive processing. Doing so, we find that two established measures, surprisal and next-word entropy, suffer from short-comings. We propose a novel measure which we name *lookahead-information gain*, that can overcome the identified problems of surprisal and next-word entropy. Next, Section 3 describes a set of recurrent neural networks that we use as PLMs, provides detail about the experimental data, and explains our statistical analysis. We then investigate how the fit of the PLM-based measures to human data relates to how accurately the PLMs have captured the linguistic patterns.

In that last step we assume that any true predictors of word processing effort should result in stronger effects when computed from increasingly more accurate PLMs. Using this method, we demonstrate that neither next-word entropy nor lookahead information gain explain human reading effort in our data.

2. Information-theoretic processing measures

PLMs assign probabilities to words (or other linguistic units) given the prior context. Formally, they estimate $P(w_t|w_1, \dots, w_{t-1})$: the probability distribution over the possibly upcoming words w_t conditional on the sequence of preceding words $w_1, \dots, w_{t-1}$. From now on, we will denote the preceding word sequence more concisely as $w_{1\dots t-1}$.

Different word prediction measures, rooted in information theory, can be derived from the probability distributions. We will discuss three: surprisal, next-word entropy, and our novel ‘lookahead information gain’ measure.

2.1. Surprisal

Hale (2001) and Levy (2008) argued that the *surprisal* of w_t (defined in Eq. (1)) is a relevant measure for the amount of cognitive processing load required to process the word.

$$S(t) = -\log P(w_t|w_{1\dots t-1}) \quad (1)$$

The surprisal measure can be derived in different ways (Levy, 2008; Smith and Levy, 2008, 2013). The following is based on Levy (2008) and highlights why surprisal has been viewed as a measure of word prediction. Suppose that the language processing system is like a perfectly predictive PLM in the sense that, after processing $w_{1\dots t-1}$ (but, crucially, before encountering any evidence about w_t) it has constructed the probability distribution $P(w_t|w_{1\dots t-1})$. That is, for each word in the language, the system has estimated the probability that it will be the upcoming word. Further, we assume that when the actual next word is encountered there is no uncertainty about its identity. This means that the probability distribution $P(w_t|w_{1\dots t-1})$ ‘collapses’ to $P(w_t|w_{1\dots t})$, where a single word has a probability of one and all other words have zero

probability.

How much cognitive work is involved in the update of the probability distribution? The so-called *Kullback-Leibler divergence* (Kullback and Leibler, 1951), also known as relative entropy or *information gain*, is an information-theoretic measure for the amount of change in a probability distribution. It is defined as:

$$D_{KL}(P||Q) = \sum_{w \in W} P(w) \log \left(\frac{P(w)}{Q(w)} \right), \quad (2)$$

where P is the updated probability distribution (in our case: $P(w_t|w_{1\dots t})$), Q is the original distribution (i.e., $P(w_t|w_{1\dots t-1})$), and W is the set of word types in the language. It is easy to prove that

$$D_{KL}(P(w_t|w_{1\dots t})||P(w_t|w_{1\dots t-1})) = S(t).$$

Hence, probabilistic prediction implies that surprisal quantifies the amount of “information gained” when processing a word, which is reflected in the amount of cognitive processing effort required for comprehension.

The validity of the surprisal measure is uncontroversial. Particularly convincing findings are that surprisal scales linearly with word-reading time (Goodkind and Bicknell, 2018; Smith and Levy, 2013) and N400 size (Yan and Jaeger), and that more accurate PLMs assign surprisal measures that correlate more strongly with reading time (Goodkind and Bicknell, 2018; Monsalve et al., 2012) and N400 size (Frank et al., 2015). However, scepticism remains about whether these results are indicative of predictive processing. After all, effects of surprisal of w_t are measured during (or after) processing the word. Hence, such effects fail to provide direct evidence that w_t was activated *before* encountering any evidence of its identity (Smith and Levy, 2013).

2.2. Next-word entropy

The *entropy* of a probability distribution is a measure of the amount of uncertainty about the outcome of the probabilistic process (Shannon, 1948). In the concrete case of word-by-word language processing, entropy thus gauges the uncertainty about the identity of the upcoming word. If there is absolute certainty about what the next word will be (i.e., $P(w_{t+1}|w_{1\dots t})$ is zero for all words except one), the entropy equals zero.³ Conversely, maximum entropy is attained when all words are equally likely to appear next. Formally, the entropy of the distribution over upcoming words is defined as:

$$H(t) = - \sum_{w_{t+1} \in W} P(w_{t+1}|w_{1\dots t}) \log P(w_{t+1}|w_{1\dots t}).$$

Because $H(t)$ is computed from the probability distribution over w_{t+1} (as opposed to w_t), effects of $H(t)$ would constitute empirical evidence for next-word prediction. However, the role of next-word entropy in language comprehension has not been investigated as thoroughly as surprisal. Roark et al. (2009) report that higher entropy over words’ parts of speech (rather than the words themselves) leads to longer self-paced reading time. More recently, van Schijndel and Linzen (2019) found the entropy over the next-word distribution to positively correlate with reading time and argued that slowdowns are caused by the readers’ increased uncertainty about the upcoming word. In an analysis of fMRI data collected while participants listened to spoken narratives, Willems et al. (2016) found that *lower* next-word entropy correlates with higher activation of, among others, the left middle frontal gyrus and right inferior frontal cortex. They interpreted this in terms of pre-activation of words when there is high certainty (i.e., low entropy) about what the

³ Note the small change in notation: we now consider the probability distribution over w_{t+1} given the words up to and including w_t . As was the case for surprisal, w_t is the word on which human processing data is compared to PLM predictions.

upcoming word could be. Lastly, Armeni et al. (2019) investigated the effect of next-word entropy on frequency-specific brain dynamics recorded using MEG, finding that increased next-word entropy coincides with higher theta-band source dynamics. Theta-band activity was maximal at middle temporal and inferior frontal lobes. Further, the authors report increased beta-band activity for decreased next-word entropy in left central and superior temporal areas, as well as in the angular gyrus.

Although these next-word entropy effects could be taken as evidence for (probabilistic) next-word prediction, there are two caveats. First, in both the fMRI data of Willems et al. (2016) and the MEG data of Armeni et al. (2019), evidence about w_{t+1} is already available at the point where the $H(t)$ effect occurs: co-articulation can provide information about the upcoming word in spoken language, for instance between determiners and nouns (Salverda et al., 2014). Moreover, the slowness of the BOLD response in MRI means that effects of $H(t)$ actually arise several words downstream.

Second, the entropy measure lacks a theoretical foundation as an estimation of processing effort. Unlike surprisal, it is not derived from assumptions about cognitive load during probabilistic processing. Why would higher next-word entropy lead to slower reading? What is the cognitive work being performed in this additional processing time?

If we attempt to derive next-word entropy similarly to surprisal (see Section 2.1) we find that this fails. We start again with the Kullback-Leibler divergence (Eq. (2)) but now, P represents the upcoming word distribution $P(w_{t+1}|w_{1...t})$ and Q is the previous distribution over w_{t+1} (i. e., before encountering w_t): $P(w_{t+1}|w_{1...t-1})$. Thus, $D_{KL}(P||Q)$ again measures the amount of probability distribution update involved in processing word w_t , but now it refers to the upcoming word w_{t+1} . Put differently, $D_{KL}(P||Q)$ quantifies the amount of cognitive work involved in generating the probabilistic next-word prediction $P(w_{t+1}|w_{1...t})$. This measure reduces to negative next-word entropy if $P(w_{t+1}|w_{1...t-1})$ is uniform, that is, the probability is divided equally over all words: $P(w_{t+1}|w_{1...t-1}) = |W|^{-1}$, where $|W|$ is the number of word types in the language. In that case:

$$D_{KL}(P||Q) = -H(t) + \log|W|,$$

where the constant term $\log|W|$ can be ignored.⁴

Clearly, this outcome is not satisfactory. First, it predicts a negative entropy effect, that is, higher $H(t)$ would correspond to faster reading of w_t . This is the opposite of what has been observed (Roark et al., 2009; van Schijndel and Linzen, 2019). Second, it is very unrealistic that all words would be considered equally likely to occur at $t+1$ after processing $w_{1...t-1}$.

2.3. Lookahead information gain

We call our novel measure *lookahead information gain* (LIG) because it quantifies the information gained (in the information-theoretic sense, see Section 2.1) due to processing w_t when the system probabilistically looks ahead to w_{t+1} . Hence, it combines aspects from the surprisal and next-word entropy measures: Like surprisal, LIG is derived from the Kullback-Leibler divergence measuring the amount of change in the probability distribution due to encountering word w_t and, like next-word entropy (but *unlike* surprisal), it is a truly forward looking measure.

We begin again with the assumption that the language processing system constructs a probability distribution $P(w_{t+1}|w_{1...t})$. The (hypothesized) amount of cognitive effort required to construct this distribution equals the Kullback-Leibler divergence from the previous distribution $P(w_{t+1}|w_{1...t-1})$.

In Section 2.2, we showed that the next-word entropy measure is

implicitly based on the unrealistic assumption that this previous distribution is uniform. However, even if the language processing system does not predict two words ahead, it could at least use the words' unconditional probabilities, that is, $P(w_{t+1}|w_{1...t-1}) = P(w)$, where $P(w)$ can simply be estimated from the word's base frequency. This reasoning yields one version of lookahead information gain:

$$LIG_1(t) = D_{KL}(P(w_{t+1}|w_{1...t})||P(w)). \quad (3)$$

If we believe that the language system predicts two words ahead (i.e., it generates $P(w_{t+1}|w_{1...t-1})$), the measure becomes:

$$LIG_2(t) = D_{KL}(P(w_{t+1}|w_{1...t})||P(w_{t+1}|w_{1...t-1})). \quad (4)$$

Studies investigating surprisal effects have shown that the cognitive effort involved in adjusting a word probability distribution is observable in online measures of reading difficulty, such as reading times and N400 sizes. Hence, in a truly predictive language processing system, we expect the LIG_1 and/or LIG_2 measures of probability-distribution update to quantify cognitive effort and therefore account for human reading data as well.

3. Methods⁵

3.1. Probabilistic language models

To estimate the information theoretic metrics of sentence processing effort outlined in Section 2 we rely on a probabilistic language model, implemented as a Long Short-Term Memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber, 1997).

For the PLMs estimating $P(w_t|w_{1...t-1})$, we use the LSTM models described in Aurnhammer and Frank (2019). The performance of a specific neural network architecture depends not only on the training materials it is presented with, but also their presentation order and on the network's random initial weights. To account for this kind of variation, Aurnhammer and Frank (2019) made use of six LSTMs, each of which was trained with different initial weights (drawn randomly from a uniform distribution between ± 0.1 ; with initial biases 0). Additionally, for each of the LSTMs the order of the training sentences, taken from the ENCOW corpus (Schäfer, 2015), was varied randomly. The entire corpus consisted of 6.47 million sentences (94,422,754 word tokens) and comprised a vocabulary of 10,103 word types.

In the subsequently described steps, all information theoretic-metrics are computed from the probability distributions averaged over the six LSTMs. We refer to this 'average language model' as PLM from here on.

For the estimation of LIG_1 we proposed a word-frequency based prior distribution $P(w_{t+1}|w_{1...t-1}) = P(w)$ (see Eq. (3)). We generate this unigram model from the word frequencies in the full training corpus. To compute LIG_2 from two fully conditional probability distributions (Eq. (4)), we use the already trained language models as forward models. After each next-word prediction step, we apply the chain rule to obtain the probability distribution over all possible words one further step ahead.⁶

In order to ascertain whether the language models were effectively trained we assess the surprisal (Eq. (1)) on the unseen experimental stimulus materials at different points during training. Better language models assign higher probabilities to the actually occurring words and thus surprisal is expected to be lower, the more sentences the PLM has been trained on. Additionally, we report on LIG_1 and LIG_2 , as specified in Eqs (3) and (4), again at different points during training in order to

⁵ All code and data will be made available at github.com/caurnhammer/AurnhammerFrankLIG

⁶ We also followed a less elegant method in which we trained six additional LSTM models to predict $P(w_{t+1})$ after seeing only $w_{1...t-1}$, that is, after observing all words up to, but not including, the current word. We provide the results of this approach in Appendix B.

⁴ See Appendix A for the derivation of this equation.

assess the behaviour of this new metric as a function of language model quality. We report on the performance of the PLM in Section 4.1.

3.2. Human sentence processing data

The information-theoretic measures we discussed above are compared to human data from three reading studies, collected using electroencephalography (EEG), eye-tracking (ET), and self-paced reading (SPR). The EEG data was published by Frank et al. (2015) and the ET and SPR data by Frank et al. (2013). The sentence stimuli used in the experiments are sampled from unpublished English novels. After half of the sentences, the participants were asked yes/no question to test comprehension.

Table 1 summarises the properties of the three experiments with regard to the number of participants, the number of word tokens and sentences used as well as the length of the sentences and the number of observations in each data set. The ET experiment made use of a subset of the SPR sentences that fitted on a single line on the computer screen. The same, shorter sentences were used in rapid serial visual presentation during the EEG experiment. In the SPR experiment, each subject received a random subset of the 361 possible sentences.

3.3. Data analysis

Our strategy to assess the fit of the information theory-based formalisations of predictive processing is to set each of the metrics in relation to the amount of training data that was presented to the language models (compare Frank et al., 2015; Goodkind and Bicknell, 2018; Monsalve et al., 2012). To this end, we compute each measure for all words in the 361 stimuli sentences from the language model after training it on 1K, 3K, 10K, 30K, 100K, 300K, 1M, 3M and 6.47M sentences. The relation of the measures to SPR times, first-pass durations, and N400 sizes is assessed for each of these training steps. At each step, the metrics of predictive processing are computed and their relation to SPR times, first-pass durations, and N400 sizes is assessed. Any observed effects should become stronger with the PLM having learned more about the language's statistical properties. Hence, if effects do not grow with the number of training sentences, we see this as reason to doubt that they truly stand in relation to the prediction effort as estimated by the PLM. Especially for less studied metrics such as next-word entropy or our new lookahead information gain, this approach is superior to individual analyses based on single fully trained models.

3.3.1. Data exclusion

In all three data sets, we exclude observations on sentence-initial and sentence-final words, words before a comma, and words with clitics. Since our reading-time analysis (Section 3.3.2) includes factors from previous words, we also remove words directly following a comma or clitic. In addition, we exclude all trials with reading times below 50 ms or over 3500 ms, and all EEG trials with artefacts (e.g., blinks) as indicated by Frank et al. (2015). Finally, we remove all non-native English speaking participants from the analyses as well as all participants who did not answer correctly to at least 80% of the yes/no comprehension questions.

Table 1

Number of participants, number of sentences, range of sentence length, average sentence length, number of tokens, and observations (after exclusion) in three human sentence reading data sets. Confer Frank et al. (2013) for details.

Exp.	Part.	Sent.	Range sent. len.	Mean sent. len.	Tokens	Obs.
EEG	24	205	5–15	9.4	1931	32,033
ET	35	205	5–15	9.4	1931	32,074
SPR	54	361	5–39	14.1	4957	136,078

3.3.2. Regression analyses

Data are analysed in linear mixed-effects regression models (LMEMs), using lme4 (Bates et al., 2015) in R (R Core Team, 2015).

Dependent variables. Reading times are log-transformed and N400 sizes, following Frank et al. (2015), are defined as the average potential on central-parietal electrodes in the 300–500 ms window respective to stimulus onset.

Independent variables. In all data sets, the regression includes as fixed effects: log-transformed word frequency in the training corpus, word length (number of characters), and word position in the sentence. Reading times are subject to so-called spillover effects, that is, processing effort at w_{t-1} affects reading time at w_t (Rayner, 1998). To account for this, we include word frequency and word length of the previous word in the analysis. The analysis of the SPR data set factors out the correlation between subsequent RTs, a typical phenomenon of the SPR paradigm, by entering the log-transformed RT on the previous word into the model. In the EEG data, baseline activity is controlled for by entering the average electrode potential in the 100 ms leading up to word onset into the regression. To account for word skipping in the eye-tracking data, a binary factor is used in the ET model, indicating whether the previous word was fixated. Additionally, the regression models include all two-way interactions between all fixed effects and a random effects structure with by-subject and by-item random intercepts, as well as by-subject random slopes for all fixed-effect predictors. All numerical predictors were scaled to z-scores.

Model comparison. To investigate the extent to which the information theoretic-metrics explain variance in reading times and N400 sizes, we employ regression model comparisons assessing the decrease in deviance. Since the significance of surprisal as a predictor in the three data sets is already well established and is a potential confound for all other formalisations of predictive processing effort, we first build a regression model that only includes the baseline factors, as described above, as well as surprisal as a fixed effect and as a by-subject random slope. For reading times, this baseline also includes surprisal on the previous word as fixed and as random effect. This model is then compared to a second model that also includes one of the other three variables (H , LIG_1 , or LIG_2), together with their previous-word counterparts for RTs (except previous word H because this equals the expected value of the current word's surprisal), again as fixed and as random effects. Mirroring this, we also obtain the unique effect of surprisal by comparing the deviance of regression models that include one of H , LIG_1 , or LIG_2 , to a second model that includes also surprisal.

The decrease in deviance between two models is expressed as a χ^2 -statistic where the degrees of freedom equals the number of additional predictors in the larger model. We indicate effects in the expected direction (i.e., positive for reading time, negative for N400 size) by positive χ^2 , and add a minus sign for reverse effects. Note that, for the negative-going N400 deflection, positive χ^2 therefore indicates stronger negativities.

4. Results

4.1. Language model training

Fig. 1 depicts the relation between the amount of training data, the D_{KL} on the word w_t (surprisal) and on the word w_{t+1} (LIG_1 and LIG_2). The better the averaged PLM becomes the higher become the probabilities assigned to the test items. As a function of this, average surprisal decreases. Unexpectedly, both LIG_1 and LIG_2 increased with training (although LIG_1 at first decreased), indicating that with better PLMs more and more probability mass needs to be shifted when updating predictions about w_t .

Fig. 1. Average Surprisal, LIG_1 , and LIG_2 of the averaged PLM at each training step, indicated by the number of training sentences.

4.2. Effects of prediction measures

Fig. 2 presents the goodness-of-fit of surprisal and next-word entropy as a function of the number of sentences the PLM was trained on.⁷ As we showed in Fig. 1, the amount of training data is a valid indication of language model quality in our models. In each of the three data sets, the effect of surprisal over and above next-word entropy is highly similar to the effect of surprisal over only the baseline factors, as reported by Aurnhammer and Frank (2019).

In the EEG data, all PLM training steps lead to next-word entropy measures that predict N400 sizes over and above what is accounted for by surprisal. However, unlike for surprisal, the LMEM comparisons indicate that more accurate next-word entropy leads to weaker regression improvements.

In the ET data, some intermediary models improved the regression models, but next-word entropy computed from the best PLM did not substantially improve the fit of the regression. The situation is even clearer in the SPR analysis in which the best PLMs lead to goodness-of-fit measures that group around $\chi^2 = 0$.

In the next series of regression models LIG_1 was analysed, leading to the goodness-of-fits depicted in Fig. 3. Entering LIG_1 values into the EEG analysis lead to model improvements with positive sign, indicating that higher LIG_1 is related to higher N400 size. However, model improvement due to LIG_1 does not increase steadily as corpus sizes of the PLM do. In the ET and the SPR data, adding LIG_1 to the regression resulted in no shape deviating from zero in clear direction. For LIG_1 there appears to be no consistent relationship between the linguistic accuracy of the PLMs and reading times.

Lastly, Fig. 4 displays model fits of LIG_2 . In the EEG data, the goodness-of-fit values obtained from lookahead information gain resemble the ones obtained from next-word entropy, with the difference that the χ^2 -values for the EEG data are larger for next-word entropy. Striking similarities can also be observed between LIG_1 and LIG_2 fits on the ET data. Resulting from the SPR model comparisons, deviations from zero- χ^2 are both positive and negative going until the final models indicate a negative relation between LIG_2 and reading time.

⁷ Networks trained on less than 10K sentences cannot have learned enough about the language's statistical properties to reliably estimate next-word probability distributions that are even slightly accurate. Nevertheless, the analyses occasionally returned unrealistically high χ^2 values ($\chi^2 > 100$) at these very early points in training. It is unclear whether these arose because of a confound, an artefact, or problem with regression model fitting, but we do not believe they can be indicative of true psycholinguistic processes. For this reason, we do not report results at 1K and 3K training sentences.

5. Discussion

In our evaluation of three information-theoretic metrics of predictive processing against three data sets of human sentence reading, each measure yielded partly non-zero results, many of which would have earned the label “significant” if they were subjected to a single-comparison null-hypothesis test. In many cases, however, these effects did not consistently grow with PLM training. This finding underlines that analysis metrics estimated by only a single PLM instance need to be interpreted with great caution. Due to the exploratory nature of our study and the large number of comparisons, we refrain entirely from claiming statistical significance. Instead we focus on assessing the relationship between language model quality (as varied by manipulating the number of training sentences) and the resulting goodness-of-fit. Based on this method none of the effects of next-word entropy and lookahead information gain can be set into a straightforward relation to the number of sentences that the PLMs were trained on. It remains to be established what leads to the striking similarities between some of the results of next-word entropy, and the two lookahead information gain variants. While the figures may suggest that, in the EEG and ET data, the effects of next-word entropy and lookahead information gain become overshadowed by the effect of surprisal for better PLMs, LMEM comparisons between regression models including next-word entropy and a baseline without surprisal take virtually the same form (see Appendix C).

5.1. Next-word entropy

Strikingly, we did not replicate the effect of next-word entropy on SPR data reported by van Schijndel and Linzen (2019). Although they used a similar amount of training data as we did in this work, their PLM was trained for many epochs and possesses two recurrent layers and is therefore superior to ours. Nevertheless, we would have expected to see an effect of next-word entropy arise over training time even with our simpler model, as it very clearly did for surprisal. Apart from the differences in PLMs, our SPR data analysis differed from van Schijndel & Linzen's in that we factored out previous-word effects of baseline variables. Further, our data set is smaller compared to the Natural Stories Corpus (Futrell et al., 2018) used by van Schijndel & Linzen which consists of 485 sentences that were read by 181 participants. These differences need further exploration to determine the disagreement between the two sets of results.

Roark et al. (2009) reported an effect on self-paced reading time of the next-part-of-speech entropy but no effect of next-word entropy. Since our analysis did not explicitly tease apart part-of-speech from word entropy, we expected our language model to indirectly learn at least some aspects of the occurrence probabilities of different parts of speech. Nonetheless, the absence of next-word entropy effects in our results is specifically clear for self-paced reading times.

5.2. The cost of predicting

In generating the lookahead information gain estimates, we observed that with better PLMs (i.e., more training) the LIG measures become larger on average. Notably this occurred regardless of whether the probability of the word after the next word was computed from base frequencies (LIG_1) or explicit prediction (LIG_2). This was unexpected given that better trained language models should gain less information from each incoming word. Why does it happen nonetheless? The Kullback-Leibler divergence, which forms the basis for our LIG definitions, can be rewritten as

$$D_{KL}(P||Q) = H(P, Q) - H(P), \quad (5)$$

where $H(P, Q)$ is the so-called cross-entropy, a measure of the difference between P and Q . For our LIG measures, P is the distribution over

Fig. 2. The goodness-of-fit (χ^2) of next-word entropy and surprisal is plotted as a function of the number of training sentences. The panels on the left, in the middle, and on the right correspond to EEG, ET, and SPR data.

Fig. 3. The goodness-of-fit (χ^2) of LIG_1 and surprisal is plotted as a function of the number of training sentences. The panels on the left, in the middle, and on the right correspond to EEG, ET, and SPR data.

upcoming words w_{t+1} after encountering the current word w_t (as estimated by the PLM), and Q is the distribution *before* encountering w_t (estimated by its unigram frequency or by using the PLM as a forward model). These two distributions become more similar over training because both PLMs learn to predict w_{t+1} . This leads to a decrease in $H(P, Q)$. However, the next-word entropy $H(P)$ decreases faster, thereby increasing D_{KL} (and hence, increasing LIG).

As is easy to see from Eq. (5), in order for LIG to decrease $H(P, Q)$ must decrease faster than $H(P)$. Apparently, this is not what happens, even for LIG_1 , which does not depend on forward modeling but uses unigram frequencies instead. Hence, our simulations show how learning to predict w_{t+1} reduces processing effort on that word but also, ironically, increases effort on w_t due to the rising cost of making more and more accurate predictions about w_{t+1} . While most theories of predictive processing claim to explain the efficiency of language processing, our approach highlights the idea that prediction also inflicts additional costs on the language comprehension system. Under strong assumptions of predictive processing, LIG would be assumed to be a useful measure of prediction that expresses the costs associated with creating novel

predictions given existing predictions. Taken to the extreme, our results may even raise doubts about the value of predictive language processing.

6. Conclusion

In this study, we replicated a well known relationship: the predictive power of surprisal on N400 sizes and reading times grows with language models that capture the linguistic structures on which they are trained increasingly well, clearly demonstrating the validity of this method. Applying this method to both next-word entropy and two variants of our newly proposed metric, lookahead information gain, we find no positive relation between any of the measures and language model quality.

If probabilistic word prediction reliably occurs across items and participants, our method and measures provide good conditions to detect potential effects of prediction effort on N400s and reading times. A thorough look at how much information is gained at each word during next-word prediction in the PLMs raises the question of how the benefits of a predictive processing mechanism relate to its costs.

Fig. 4. The goodness-of-fit (χ^2) of LIG_2 and surprisal is plotted as a function of the number of training sentences. The panels on the left, in the middle, and on the right correspond to EEG, ET, and SPR data.

Taken at face value our study does not support predictive processing. This holds at least to the extent to which our word-prediction model captures the information that humans would presumably capitalise on for their predictions. In this regard it is worth noting that the language model approximations of predictive processing employed in this study are by definition incomplete. They are based only on the *words* present

in *one sentence* up until and including the currently processed word. As such, the PLM does not account for many other predictive cues, such as extra-sentential discourse context (Otten and van Berkum, 2008), world knowledge (Bicknell et al., 2010), non-verbal factors (Rommers et al., 2015).

Appendix A. Derivation of next-word entropy

In Section 2.3, we claimed that the Kullback-Leibler divergence between the previous and current probability distribution over w_{t+1} equals the negative next-word entropy if the previous distribution is assumed to be uniform, that is, $P(w_{t+1}) = |W|^{-1}$. This is derived as follows:

$$\begin{aligned}
 D_{KL}(P||Q) &= D_{KL}\left(P(w_{t+1}|w_{1...t}) || |W|^{-1}\right) \\
 &= \sum_{w \in W} P(w_{t+1}|w_{1...t}) \log\left(\frac{P(w_{t+1}|w_{1...t})}{|W|^{-1}}\right) \\
 &= \sum_{w \in W} P(w_{t+1}|w_{1...t}) (\log P(w_{t+1}|w_{1...t}) + \log |W|) \\
 &= \sum_{w \in W} P(w_{t+1}|w_{1...t}) \log P(w_{t+1}|w_{1...t}) + \log |W| \sum_{w \in W} P(w_{t+1}|w_{1...t}) \\
 &= -H(t) + \log |W|
 \end{aligned}$$

Appendix B. Computing LIG_2 from two separate PLMs

As an alternative to forward modeling, the probabilities of w_{t+1} can be estimated by a language model that estimates w_{t+1} after seeing only the words in a sentence up until (and including) w_{t-1} . Using identical network architectures and training regimes as described in 3.1, we computed LIG_2 by taking $P(w_{t+1}|w_{1...t})$ from the LSTMs used before but $P(w_{t+1}|w_{1...t-1})$ from LSTMs predicting two words ahead. The results of this approach are provided in Figure B.5.

Fig. B.5. The goodness-of-fit (χ^2) of LIG_2 computed from a PLM predicting the next word and a PLM predicting two words ahead. Surprisal is plotted as a function of the number of training sentences. The panels on the left, in the middle, and on the right correspond to EEG, ET, and SPR data.

Appendix C. Individual effects of prediction measures

Figures C.6, C.7, C.8, and C.9 provide the goodness-of-fit of surprisal, H , LIG_1 , and LIG_2 , respectively, over a regression model that includes only baseline effects.

Fig. C.6. Goodness-of-fit of surprisal over baseline factors.

Fig. C.7. Goodness-of-fit of next-word entropy over baseline factors.

Fig. C.8. Goodness-of-fit of LIG_1 over baseline factors.

Fig. C.9. Goodness-of-fit of LIG_2 over baseline factors.

Funding

The research 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.

Declaration of interests

None.

References

- Armeni, K., Willems, R.M., van den Bosch, A., Schoffelen, J.-M., 2019. Frequency-specific brain dynamics related to prediction during language comprehension. *Neuroimage* 198, 283–295.
- Aurnhammer, C., Frank, S.L., 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In: *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX, pp. 112–118.
- Bar, M., 2007. The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci.* 11, 280–289.
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48.
- Bicknell, K., Elman, J.L., Hare, M., McRae, K., Kutas, M., 2010. Effects of event knowledge in processing verbal arguments. *J. Mem. Lang.* 63, 489–505.
- Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2017. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803–809.
- Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204.
- Corps, R.E., Gambi, C., Pickering, M.J., 2018. Coordinating utterances during turn-taking: the role of prediction, response preparation, and articulation. *Discourse Process* 55, 230–240.
- Dambacher, M., Rolfs, M., Göllner, K., Kliegl, R., Jacobs, A.M., 2009. Event-related potentials reveal rapid verification of predicted visual input. *PLoS One* 4, e5047.
- DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121.
- den Ouden, H.E., Kok, P., de Lange, F.P., 2012. How prediction errors shape perception, attention, and motivation. *Front. Psychol.* 3.
- Dikker, S., Pykkänen, L., 2013. Predicting language: MEG evidence for lexical preactivation. *Brain Lang.* 127, 55–64.
- Dikker, S., Rabagliati, H., Farmer, T.A., Pykkänen, L., 2010. Early occipital sensitivity to syntactic category is based on form typicality. *Psychol. Sci.* 21, 629–634.
- Federmeier, K.D., 2007. Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505.
- Frank, S.L., Monsalve, I.F., Thompson, R.L., Vigliocco, G., 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behav. Res. Methods* 45, 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 Lang.* 140, 1–11.
- Friston, K., 2010. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138.
- Futrell, R., Gibson, E., Tily, H.J., Blank, I., Vishnevetsky, A., Piantadosi, S., Fedorenko, E., 2018. The natural Stories corpus. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association, Paris, France, pp. 76–82.
- 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*. Association for Computational Linguistics, Stroudsburg, PA, pp. 10–18.
- Hahn, M., Keller, F., 2016. Modeling human reading with neural attention. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, pp. 85–95.
- Hale, J., Lutz, D., Luh, W.-M., Brennan, J., 2015. Modeling fMRI time courses with linguistic structure at various grain sizes. In: *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, pp. 89–97.
- 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*. Association for Computational Linguistics, Stroudsburg, PA, pp. 159–166.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- Huetting, F., 2015. Four central questions about prediction in language processing. *Brain Res.* 1626, 118–135.
- Jackendoff, R., 2007. A parallel architecture perspective on language processing. *Brain Res.* 1146, 2–22.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.
- Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647.
- Laszlo, S., Federmeier, K.D., 2009. A beautiful day in the neighborhood: an event-related potential study of lexical relationships and prediction in context. *J. Mem. Lang.* 61, 326–338.
- Lau, E.F., Weber, K., Gramfort, A., Hämäläinen, M.S., Kuperberg, G.R., 2014. Spatiotemporal signatures of lexical–semantic prediction. *Cerebr. Cortex* 26, 1377–1387.
- Levy, R., 2008. Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177.
- 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*. Association for Computational Linguistics, Stroudsburg, PA, pp. 398–408.
- Nicenboim, B., Vasishth, S., Rösler, F., 2019. Are Words Pre-activated Probabilistically during Sentence Comprehension? Evidence from New Data and a Bayesian Random-Effects Meta-Analysis Using Publicly Available Data. *PsyArXiv Preprint*. <https://doi.org/10.31234/osf.io/2atrh>.
- Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthum, S.v., Bartolozzi, F., Kogan, V., Ito, A., et al., 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7.
- Otten, M., van Berkum, J.J., 2008. Discourse-based word anticipation during language processing: prediction or priming? *Discourse Process* 45, 464–496.
- R Core Team, 2015. R: A Language and Environment for Statistical Computing.
- Rabovsky, M., 2019. Implicit Semantic Prediction Error Can Account for N400 Effects on Articles that Do Not Differ in Meaning: A Neural Network Model. *BioRxiv Preprint*. <https://doi.org/10.1101/569020>.
- Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422.
- Roark, B., Bachrach, A., Cardenas, C., Pallier, C., 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, pp. 324–333.
- Rommers, J., Meyer, A.S., Huettig, F., 2015. Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Atten. Percept. Psychophys.* 77, 720–730.
- Salverda, A.P., Kleinschmidt, D., Tanenhaus, M.K., 2014. Immediate effects of anticipatory coarticulation in spoken-word recognition. *J. Mem. Lang.* 71, 145–163.
- Schäfer, R., 2015. Processing and querying large web corpora with the COW14 architecture. In: *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora*. Institut für Deutsche Sprache, Mannheim, Germany, pp. 28–34.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 623–656.
- Smith, N.J., Levy, R., 2008. Optimal processing times in reading: a formal model and empirical investigation. In: *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, Austin, TX, pp. 595–600.
- Smith, N.J., Levy, R., 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319.
- Staub, A., 2015. The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. *Language and Linguistics Compass* 9, 311–327.
- Staub, A., Clifton, J., Charles, 2006. Syntactic prediction in language comprehension: evidence from either... *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 425–436.
- van Berkum, J.J., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467.
- van Petten, C., Luka, B.J., 2012. Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* 83, 176–190.
- van Schijndel, M., Linzen, T., 2019. Can entropy explain successor surprisal effects in reading? In: *Proceedings of the Society for Computation in Linguistics*.
- Wicha, N.Y., Moreno, E.M., Kutas, M., 2004. Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16, 1272–1288.
- Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., van den Bosch, A., 2016. Prediction during natural language comprehension. *Cerebr. Cortex* 26, 2506–2516.
- Yan, S., & Jaeger, T. F. (in press). (Early) context effects on event-related potentials over natural inputs. *Language, Cognition and Neuroscience*, .