Brain & Language 140 (2015) 1–11

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

The ERP response to the amount of information conveyed by words in sentences

Stefan L. Frank [a,b,*], Leun J. Otten [c], Giulia Galli [c,d], Gabriella Vigliocco [a]

[a] *Department of Experimental Psychology, University College London, United Kingdom*
[b] *Centre for Language Studies, Radboud University Nijmegen, Netherlands*
[c] *Institute of Cognitive Neuroscience, University College London, United Kingdom*
[d] *Department of Medicine, Surgery and Neuroscience, University of Siena, Italy*

ARTICLE INFO

Article history:
Accepted 18 October 2014
Available online 17 November 2014

Keywords:
Information theory
Surprisal
Entropy
Event-related potentials
Sentence comprehension
Reading

ABSTRACT

Reading times on words in a sentence depend on the amount of information the words convey, which can be estimated by probabilistic language models. We investigate whether event-related potentials (ERPs), too, are predicted by information measures. Three types of language models estimated four different information measures on each word of a sample of English sentences. Six different ERP deflections were extracted from the EEG signal of participants reading the same sentences. A comparison between the information measures and ERPs revealed a reliable correlation between N400 amplitude and word surprisal. Language models that make no use of syntactic structure fitted the data better than did a phrase-structure grammar, which did not account for unique variance in N400 amplitude. These findings suggest that different information measures quantify cognitively different processes and that readers do not make use of a sentence's hierarchical structure for generating expectations about the upcoming word.

1. Introduction

Much recent computational work in psycholinguistics has called upon insights from information theory to bridge between psycholinguistic experiments and statistical models of language. Jaeger (2010), for example, argues that information-theoretic considerations can explain speakers' structural choices in sentence production. Likewise, in sentence comprehension, each word conveys a certain amount of information and – to the extent that language comprehension is information processing – this amount should be predictive of how much cognitive effort is required to process the word (Hale, 2006; Levy, 2008). The amount of information conveyed by a word (or *word information* for short) can be computed from probabilistic models of the language, whereas the amount of cognitive effort involved in processing a word can be observed, for example by measuring word reading times. Comparisons between word-information values and reading times have indeed revealed that more informative words take longer to read (e.g., Frank, 2013; Smith & Levy, 2013).

Studies that investigate how word information relates to reading time are not necessarily concerned with explaining any particular psycholinguistic phenomenon. Rather, they tend to apply large-scale regression analyses to uncover the general relation between quantitative predictions and reading times on each word of a text corpus. In the current paper, we apply such a parametric (non-factorial) experimental design to investigate the effect of word information on the ERP response during sentence reading. That is, we bridge between computational, probabilistic models of language processing and the neural computations involved in sentence comprehension.

1.1. Quantifying word information

The rapid serial visual presentation procedure that is typical for EEG reading studies (and was also applied in our experiment) enforces that all words are read in strictly serial order. Hence, the comprehension process for a k-word sentence can be assumed to comprise a sequence of comprehension events for k words: $w_1, w_2, \ldots, w_k$, or $w_{1\ldots k}$ for short. The different measures of information that have been put forth as cognitively relevant to sentence processing are all rooted in a probabilistic formalization of such word-by-word comprehension.

After the first t words of the sentence, $w_{1\ldots t}$, have been processed, the identity of the upcoming word, w_{t+1}, is still unknown and can therefore be viewed as a random variable. The *surprisal* (or 'self information') of the outcome of a random variable is

[*] Corresponding author at: Centre for Language Studies, Radboud University Nijmegen, Netherlands.
E-mail address: s.frank@ucl.ac.uk (S.L. Frank).

http://dx.doi.org/10.1016/j.bandl.2014.10.006
0093-934X/© 2014 The Authors. Published by Elsevier Inc.

defined as the negative logarithm of the outcome's probability, which in this case is the probability of the actual next word w_{t+1} given the sentence so far:

$$\text{surprisal }(w_{t+1}) = -\log P(w_{t+1}|w_{1\ldots t}), \tag{1}$$

where the base of the logarithm forms an arbitrary scaling factor (we use base-e). Informally, the surprisal of a word can be viewed as a measure of the extent to which its occurrence was unexpected.

The symbols w in Eq. (1) do not need to stand for actual words. Instead, they may represent the words' syntactic categories (i.e., their parts-of-speech; PoS), in which case Eq. (1) formalizes the unexpectedness of the encountered PoS given the PoS-sequence corresponding to the sentence so far. This does away with any (lexical) semantics and may thereby reveal purely syntactic effects (cf. Frank & Bod, 2011).

Several authors have put forth theoretical arguments for surprisal as a measure of cognitive processing effort or predictor of word reading time (Hale, 2001; Levy, 2008; Smith & Levy, 2008; Smith & Levy, 2013) and it is indeed well established by now that reading times correlate positively with the surprisal of words (Fernandez Monsalve, Frank, & Vigliocco, 2012; Fossum & Levy, 2012; Frank, 2014; Frank & Thompson, 2012; Mitchell, Lapata, Demberg, & Keller, 2010; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2013) as well as with the surprisal of parts-of-speech (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008; Frank & Bod, 2011).

A second important concept from information theory is *entropy* (Shannon, 1948), a measure of the uncertainty about the outcome of a random variable. For example, after processing $w_{1\ldots t}$, the uncertainty about the remainder of the sentence is quantified by the entropy of the distribution of probabilities over the possible continuations $w_{t+1\ldots k}$ (with $k > t$). This entropy is defined as

$$H(W_{t+1\ldots k}) = -\sum_{w_{t+1\ldots k}} P(w_{t+1\ldots k}|w_{1\ldots t}) \log P(w_{t+1\ldots k}|w_{1\ldots t}), \tag{2}$$

where $W_{t+1\ldots k}$ is a random variable with the particular sentence continuations $w_{t+1\ldots k}$ as its possible outcomes. When the next word or part-of-speech, w_{t+1}, is encountered, this will usually decrease the uncertainty about the rest of the sentence, that is, $H(W_{t+2\ldots k})$ is generally smaller than $H(W_{t+1\ldots k})$. The difference between the two is the *entropy reduction*, which will be denoted ΔH. Entropy is strongly reduced when moving from a situation in which there exists many possible, low-probability continuations to one in which there are few, high-probability continuations. Informally, entropy reduction can be said to quantify how much ambiguity is resolved by the current word or PoS, at least, to the extent that disambiguation reduces the number of possible sentence continuations.

Hale (2003, 2006, 2011) argues that entropy reduction quantifies the amount of cognitive processing effort during sentence comprehension. However, ΔH as defined here is a simplification of Hale's original proposal, which relies on syntactic structures rather than mere word strings (see Frank, 2013). Reading times are indeed predicted by ΔH, both when defined over words (Frank, 2013) and over parts-of-speech (Frank, 2010), even after factoring out surprisal. Another variation of entropy reduction has also been shown to correlate with reading times (Wu, Bachrach, Cardenas, & Schuler, 2010).

To summarize, we use four definitions of the amount of information conveyed: the surprisal of words or their PoS, and the entropy reduction due to words or their PoS.

1.2. The present study

Our current objectives are twofold. First, we wish to investigate whether a relation between word information and ERP amplitude indeed exists. We looked at six different ERP components, three of which are generally viewed as indicative of lexical, semantic, or conceptual processing; these are the N400, and (Early) Post-N400 Positivity (EPNP and PNP) components. The other three have been claimed to reflect syntactic processing: (Early) Left Anterior Negativity (ELAN and LAN) and P600. Because we have defined information not only for each word in a sentence but also for the word's syntactic category, ERP components that are related to either lexical or syntactic processing can potentially be distinguished. Likewise, we compare the surprisal and entropy reduction measures. In particular, an effect of word surprisal is expected on the size of the N400, a negative-going deflection with a centro-parietal distribution, peaking at about 400 ms after word onset. Previous work (Dambacher, Kliegl, Hofmann, & Jacobs, 2006) has shown that this component correlates with cloze probability, which can be taken as an informal estimate of word probability, based on human judgments rather than statistical models. In addition, Parviz, Johnson, Johnson, and Brock (2011) estimated surprisal on sentence-final nouns appearing in either low- or high-constraining sentence context that made the nouns less or more predictable. They found that the N400 (as measured by MEG) was sensitive to surprisal. However, no effect of surprisal remained after factoring out context constraint.

It is much harder to derive clear predictions for the other ERP components and alternative notions of word information. We return to this issue in Section 4.2, which discusses why relations between particular information measures and ERP components may be expected on the basis of the current literature.

Second, the use of model-derived rather than cloze probabilities allows us to compare the explanatory value of different probabilistic language models. Any such model can estimate the probabilities required to compute surprisal and entropy, at least in principle. However, models differ in their underlying assumptions about the linguistic structures and mechanisms involved in sentence comprehension. A model whose assumptions are closer to cognitive reality should give rise to information measures that are more predictive of experimental data. Hence, the most plausible cognitive mechanisms for sentence processing can be identified by comparing different models' abilities to explain the ERPs. This approach to selection among sentence comprehension models has previously been applied successfully using reading time data from eye tracking studies (Frank & Bod, 2011; Frank & Thompson, 2012). Here, we compare three model types that are based on very different assumption: *n*-gram models, which do not embody any cognitive or linguistic theory; recurrent neural networks, which are domain-general temporal learning and processing systems; and phrase-structure grammars, which capture hierarchical syntactic structure.

2. Methods

2.1. EEG data collection

2.1.1. Participants

Twenty-four healthy, adult volunteers (10 female, mean age 28.0 years) from the UCL Psychology subject pool took part in the reading study. All were right handed and native speakers of English. They were paid £15 for their participation.

2.1.2. Materials

As the current study aimed at investigating the general relation between word information and ERP amplitudes, the sentence stimuli were not intended to manipulate any particular linguistic construction or psychological factor. Rather, they were sampled to be representative of written British English. The use of naturally occurring materials rather than hand-crafted experimental stimuli increases the generalizability of results.

We took the 205 sentences (comprising 1931 word tokens) from the UCL corpus of reading times (Frank, Fernandez Monsalve, Thompson, & Vigliocco, 2013) for which eye-tracking data are available. These sentences, which came from three little known novels, do not contain any syntactic violations, semantic anomalies, or other unnatural use of language. One hundred and ten (54%) of the sentences were paired with a yes/no comprehension question to ensure that participants read attentively. For further details, including the list of stimuli, see Frank et al. (2013).

2.1.3. Procedure

The sentences were presented in random order. Each sentence's presentation was preceded by a centrally located fixation cross. As soon as the participant pressed a key, the cross was replaced by the sentence's first word, which was then automatically replaced by each subsequent word. Words were always centrally located on the monitor, printed in 24-point Courier New font, in black letters on a 10% gray background. Word presentation duration (ignoring the variable delay caused by the screen refresh rate) equalled $190 + 20m$ ms, where m is the number of characters in the word, including any attached punctuation. Such word-length dependent presentation duration allows for more natural reading compared to a fixed presentation rate (Nieuwland & Van Berkum, 2006). After the word disappeared, there was a 390 ms interval before appearance of the next word, making the theoretically shortest SOA (i.e., following one-letter words) equal to $390 + 190 + 20 = 600$ ms. In reality, the screen-refresh delay yielded a minimum SOA of 627 ms (mean: 700 ms; SD: 34 ms). A technical malfunction resulted in much longer than intended SOA at three occasions. Data on the corresponding sentences (one for each of three subjects) was not analyzed.

The comprehension question (if any) was presented directly after offset of the sentence-final word. The next sentence's fixation cross appeared as soon as the subject answered the question, or after key press if there was no question. All participants answered at least 80% of the comprehension questions correctly.

Participants were urged to minimize blinks, eye movements, and head movements during sentence presentation. They were encouraged to take a few minutes break after reading 50, 100, and 150 sentences. A complete session, including fitting of the EEG cap, took approximately 1.5 h.

2.1.4. EEG recording and preprocessing

The EEG signal was recorded continuously at a rate of 500 Hz from 32 scalp sites (montage M10, see Fig. 3 and www.easycap.de) and the two mastoids relative to a midfrontal site using silver/silver-chloride electrodes with impedances below 5 kΩ. Vertical eye movements were recorded bipolarly from electrodes above and below the right eye, and horizontal eye movements from electrodes at the outer canthi. Signals were band-pass filtered online between 0.01 and 35 Hz. Offline, signals were filtered between 0.05 and 25 Hz (zero phase shift, 96 dB roll-off), downsampled to 250 Hz, and re-referenced to the average of the two mastoids, reinstating the frontal electrode site.

The signal was epoched into trials ranging from 100 ms before until 924 ms after each word onset. Any trial with a peak amplitude of over 100 μV was removed. Further artifacts (mostly due to eye blinks) were identified by visual inspection and corresponding trials were removed.

2.2. Estimating word information

2.2.1. Training corpus

The conditional probabilities in Eqs. (1) and (2), required to compute surprisal and entropy, can be accurately estimated by any probabilistic language model that is trained on a large text corpus. Our corpus consisted of 1.06 million sentences from the written-text part of the British National Corpus (BNC), selected by taking the 10,000 most frequent word types from the full BNC and then extracting all BNC sentences that contain only those words. The corresponding parts-of-speech were obtained by applying the Stanford parser (Klein & Manning, 2003) to the selected BNC sentences, resulting in syntactic tree structures where each word token is assigned one of 45 PoS labels (following the Penn Treebank PoS-tagging guidelines; Santorini, 1991).

2.2.2. Language models

We applied three model types that vary greatly in their underlying assumptions: n-gram models (also known as Markov models), recurrent neural networks (RNNs), and probabilistic phrase-structure grammars (PSGs). An n-gram model estimates the probability of a word by taking only the previous $n-1$ words into account. That is, $P(w_{t+1}|w_{1\ldots t})$ is reduced to $P(w_{t+1}|w_{t-n+2\ldots t})$. Because of this radical simplification, n-gram models are not considered cognitively or linguistically realistic. Nevertheless, they can be remarkably accurate because the n-gram probabilities can be estimated efficiently and accurately by simply counting the frequencies of very short words strings $w_{t-n+2\ldots t}$ and $w_{t-n+2\ldots t+1}$ in the training corpus.

The SRILM software (Stolcke, 2002) was used to train three n-gram models (with $n = 2$, 3, and 4) on the 1.06 million selected BNC sentences, using modified Kneser–Ney smoothing (Chen & Goodman, 1999). Three more models (with $n = 2$, 3, and 4) were trained on the sentences' PoS. The simplicity of n-gram models makes it feasible to train them on very large data sets, so three additional models (again with $n = 2,$ 3, and 4) were obtained by training on the 4.8 million sentences of the full BNC.

The RNN is like an n-gram model in the sense that it is trained on unanalyzed word sequences rather than syntactic structures. However, it is sensitive to all of the sentence's previous words, and not just the previous $n-1$, because it uses an internal layer of units to integrate over the entire word sequence. It does so by combining the input representing the current word w_t with the current state of the internal layer, which itself depends on the entire sequence of previous inputs $w_{1\ldots t-1}$ (see Elman, 1990). Such systems have been widely applied to cognitive modeling of temporal processing, also outside the linguistic domain, because (unlike the PSG model) they do not rely on any particular linguistic assumption. For example, they do not assume syntactic categories or hierarchical structure.

The RNN model was identical in both architecture and training procedure to the one presented by Fernandez Monsalve et al. (2012) and Frank (2013), except that the current RNN received a larger number of word types and sentences for training. Its output after processing the sentence-so-far $w_{1\ldots t}$ is a probability distribution $P(w_{t+1}|w_{1\ldots t})$ over all word types. That is, at each point in a sentence, the network estimates the probability of each possible upcoming word.

The number of different parts-of-speech is much smaller than the number of word types (45 versus 10,000). Consequently, a much simpler RNN architecture (Elman's, 1990, simple recurrent network) suffices for modeling PoS-sequences.

To obtain a range of increasingly accurate models, nine training corpora of different sizes were constructed by taking increasingly large subsets of the training sentences, such that the smallest subset held just 2000 sentences and largest contained all 1.06 million. The networks were trained on each of these, as well as on all 1.06 million BNC sentences twice, yielding a total of ten RNN models trained on words and ten trained on parts-of-speech.

Unlike n-gram and RNN models, PSGs are based on linguistic insights about the hierarchical syntactic structure of sentences. A probabilistic grammar assigns possible structures to a sentence, as well as probabilities to the structures. From these follow the probabilities of the sentence's words.

The training corpus for the PSGs was the set of selected BNC sentences' syntactic structures, as assigned by the Stanford parser. A PSG was extracted from each of the nine, incrementally large subsets of the selected BNC sentences (as explained above)[1] by Roark's (2001) PSG-induction algorithm. Nine PSGs defined over PoS-strings were obtained by the same procedure, except that the words were removed from the training sentences' syntactic structures, leaving the parts-of-speech to play the role of words.

2.2.3. Surprisal and linguistic accuracy

After training, the language models were presented with the same 205 sentences as read by the participants in our EEG study. Generating surprisal values for these sentences is straightforward because all three model types directly output a probability estimate for each word.

A particular model's surprisal estimates also serve to quantify how well that model has captured the statistical patterns of English sentences: Good language models form accurate expectations about the upcoming words so generally assign high probability (i.e., low surprisal) to words that actually appear. Hence, we take the average log-transformed word probability over the experimental sentences as a measure of a model's *linguistic accuracy* (Frank & Bod, 2011).[2] Although this measure says nothing about the model's ability to account for ERP data, we would expect models with higher linguistic accuracy to provide better fit to the ERP amplitudes because such models more closely capture the linguistic knowledge of our native English speaking participants.

2.2.4. Entropy reduction

The word-sequence probabilities required for computing entropy (Eq. (2)) follow from the next-word probabilities by application of the chain rule: $P(w_{t+1\ldots k}|w_{1\ldots t}) = \prod_{i=1}^{k} P(w_{t+i}|w_{1\ldots t+i-1})$. However, the number of word sequences grows exponentially with sequence length, resulting in a combinatorial explosion when attempting to compute all the $P(w_{t+1\ldots k}|w_{1\ldots t})$ for anything but very short sequences $w_{t+1\ldots k}$. The RNN model fares better in this respect than the other two model types because it computes the probability distribution $P(w_{t+1}|w_{1\ldots t})$ over all word types in parallel. This distribution can be fed back as input into the network to get the distribution at $t+2$, etc. For this reason, only the RNN model was used to estimate entropy. Following Frank (2013), the computation was simplified by retaining only the 40 most probable word sequences when feeding back the probability distribution (no such restriction applied to the computation of PoS entropy). Furthermore, the 'lookahead distance' was restricted to $k \leqslant 4$, that is, no more than four upcoming words or PoS (i.e., sequences $w_{t+1\ldots t+4}$, or shorter) are taken into account when computing entropy.

It has been shown that this RNN-based simplified entropy reduction measure suffices to explain variance in word reading times over and above what is already explained by surprisal. However, it strongly depends on the value of k: In an analysis of ΔH-values defined over words, Frank (2013) found that larger k resulted in stronger correlation with reading time, reaching statistical significance when $k > 2$.

2.3. Data analysis

2.3.1. ERP components

Six ERP components of interest were chosen on the basis of the literature on ERP studies using visually presented sentences. Table 1 shows the time window (relative to word onset) and sites assigned to each component, as well as references to the studies on which these assignments were based. Because of differences in EEG cap montage, some of the selected electrode locations only approximated those from the cited studies. Also, the time window of the PNP component was reduced to 600–700 ms (from Thornhill and Van Petten's 600–900 ms) so that the PNP resulting from the current word is only minimally (if at all) affected by the upcoming word that can appear as soon as 627 ms after the current word's onset. The ERP amplitude for a particular component, subject, and word token was defined as the average scalp potential over the ERP's time window and electrode sites as listed in Table 1.

Our interest in ERP effects at each word, in combination with the uncontrolled nature of the stimuli, makes it difficult to prevent large differences in EEG baselines. Simply subtracting baseline ERPs from the amplitudes can cause artifacts, in particular for early components (see, e.g., Steinhauer & Drury, 2012). One safe and efficient method for mitigating the baseline problem is to reduce the correlation between the ERP baselines and amplitudes by applying an additional high-pass filter with a sufficiently high cut-off frequency. We compared the correlations between ERP baselines (determined by averaging over each component's electrodes in the 100 ms leading up to word onset) and amplitudes after applying 0.25 Hz, 0.33 Hz, or 0.50 Hz high-pass filters,[3] or no additional filter. As can be seen in the online supplementary materials, the 0.50 Hz filter yielded the weakest correlation overall, so this filter was used to compute the amplitudes for subsequent data analysis.

Our statistical analyses assume normally distributed data, but the distribution of amplitudes was far from normal for the ELAN, LAN, EPNP, and PNP components: Their excess kurtosis ranged from +1.33 to +6.21 where values between ±1 are generally considered acceptable. Therefore, the modulus transformation (John & Draper, 1980) was applied to these components, bringing all excess kurtosis values below 1. All six ERP amplitude distributions were nearly symmetrical (skewness was between -0.149 and $+0.025$) so their divergence from normality is negligible.

2.3.2. Quantifying fit to ERP amplitude

Words attached to a comma, clitics, sentence-initial, and sentence-final words were discarded from further analysis, leaving a grand total of 31,997 analyzed data points per investigated ERP component.

The ERP amplitudes were not averaged over subjects or items. Instead, variance among subjects and among items is taken into account by fitting a linear mixed-effects regression model to each set of ERP amplitudes (the same approach was applied by Dambacher et al., 2006). These regression models included as standardized covariates: log-transformed word frequency, word length (number of characters), word position in the sentence, sentence position in the experiment, and all two-way interactions between these. In addition, there were by-subject and by-item random intervals, as well as the maximal by-subject random slope structure (as advocated by Barr, Levy, Scheepers, & Tilly, 2013).

[1] The smallest subset held 3000 rather than 2000 sentences, because not all experimental sentences could be parsed by a grammar trained on the 2000-sentence set.

[2] To be precise, we take the average over all words in the experimental sentences, weighted by the number of times the word takes part in the analysis of the ERP data. The models' linguistic accuracies are presented in the supplementary materials.

[3] The filters were applied to the continuous EEG signal during each sentence presentation, from 100 ms before the first word onset up to 924 ms after the sentence-final word onset. The rest of the signal was not included because participants were encouraged to blink and move between sentence presentations, causing many artifacts. Since a sentence's first and last word are not included in the analysis, there is no risk of edge-effects from filtering each sentence individually.

Table 1
Definitions of ERP components: time windows and electrode site numbers (montage M10). See Fig. 3 for approximate scalp locations.

Name	Time window (ms)	Electrode sites	Reference
ELAN	125–175	8, 21, 22, 33, 34, 37, 49,	Gunter et al. (1999)
LAN	300–400	8, 18, 33, 34, 48, 49	Kaan and Swaab (2003a)
N400	300–500	1, 14, 24, 25, 26, 29, 30, 31, 41, 42, 44, 45	Dambacher et al. (2006)
EPNP	400–600	35, 36, 37, 49, 50	Thornhill and Van Petten (2012)
P600	500–700	1, 12, 14, 16, 24, 25, 26, 29, 30, 31, 39, 40, 41, 42, 44, 45, 46, 47	Van Berkum et al. (2007)
PNP	600–700	1, 8, 10, 18, 21, 22, 24, 31, 33, 34	Thornhill and Van Petten (2012)

As mentioned above, no baseline correction was applied because of the risk of introducing artifacts. Instead, ERP baseline is also included as a factor in the regression model. This factors out any systematic difference in ERP amplitude that is already present pre-stimulus, whereas no post-stimulus 'effects' can be artificially introduced.

The regression models so far do not include a factor for word information. When including as a predictor the estimates of word surprisal under a particular language model, the regression model's deviance decreases. The size of this decrease is the χ^2-statistic of a likelihood-ratio test for significance of the surprisal effect and is taken as the measure of the fit of surprisal to the ERP amplitudes. This definition equals what Frank and Bod (2011) call 'psychological accuracy' in an analysis of reading times. The same method is applied for obtaining measures for quantifying the fit of entropy reduction and PoS surprisal, with one caveat: The regression models already include a factor for word surprisal (estimated by the 4-gram model trained on the full BNC because this model had the highest linguistic accuracy). Consequently, the χ^2 measures for entropy reduction and PoS surprisal quantify their fit over and above what is already explained by word surprisal.

2.3.3. Exploratory and confirmatory analyses

We have no strong expectations about which information measure correlates with which ERP component, apart from the relation between word surprisal and the N400. Therefore, the current study is mostly exploratory, which means that it suitable for *generating* hypotheses but not for *testing* them (cf. De Groot, 2014). Strictly speaking, conclusions can only be drawn after a subsequent confirmatory study with new data. To be able to draw conclusions from our data, we divide the full data set into two subsets: the Exploratory Data, comprising only the 12 odd-numbered subjects; and the Confirmatory Data, comprising the 12 even-numbered subjects. The Exploratory Data is used to identify the information measures and ERP components that are potentially related. Only these potential effects are then tested on the Confirmatory Data. As potential effects, we consider only the ones for which all of the following conditions hold:

1. At least one of the language models results in a $\chi^2 > 3.84$ (the critical value at the $\alpha = .05$ level).
2. The direction matches the ERP component, that is, larger information value corresponds to more negative-going (E)LAN/N400 and to more positive-going (E)PNP/P600.
3. More accurate language models result in stronger effects, as apparent in larger χ^2.
4. For entropy reduction: larger lookahead distance k results in stronger effects, as apparent in larger χ^2.

3. Results

3.1. Exploratory and confirmatory analyses

As displayed in Fig. 1, the exploratory analysis identified four potential effects: Word surprisal seems to predict the amplitude of N400 and, to a much lesser extent, LAN; Word entropy reduction may explain EPNP and, to a much lesser extent, PNP. There are no potential effects of the PoS information measures (see the supplementary materials for all exploratory results).

Of the four potential effects, only the N400 survives in the Confirmatory Data (see Fig. 2). All model types reach $\chi^2 > 11$ for this component, which corresponds to $p < .001$. Hence, we have reliable evidence for an effect of word surprisal on the N400 but not for any other relation between word (or PoS) information and any ERP component.

Having established that a word surprisal effect occurs in both the Exploratory and Confirmatory Data sets, we now take the full set of data to investigate whether the effect can indeed be considered an N400. To this aim, Fig. 3 plots average ERP wave forms at each electrode, separately for words with low (bottom third) and high (top third) word surprisal as estimated by the 4-gram model because this model showed the strongest overall effect on the N400 (see Fig. 4). The high-surprisal words result in a more negative deflection than the low-surprisal words, in particular within the 300–500 ms time window and at central sites, as is typical for the N400. Hence, word surprisal indeed affects N400 amplitude. The corresponding regression coefficient ranges from −0.17 (for the *n*-gram model) to −0.22 (for RNN), which is to say that one standard deviation increase in surprisal corresponds to an average increase in N400 amplitude of between 0.17 and 0.22 µV.

3.2. Comparing word classes

Because nearly all studies that find N400 effects are concerned with content words only, it is of interest to perform separate

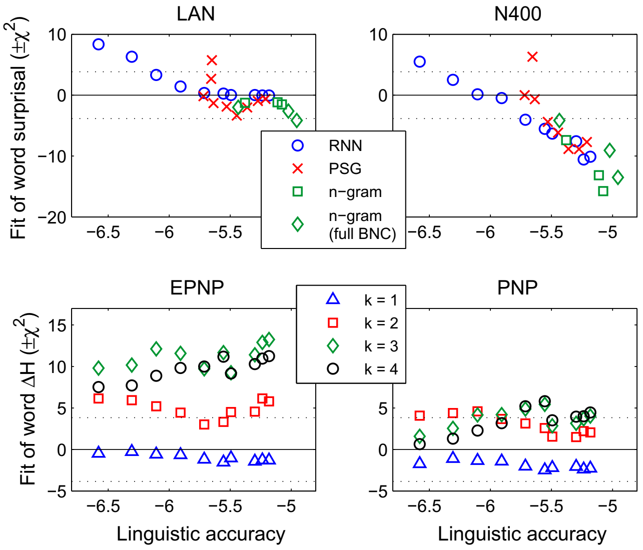

Fig. 1. Potential effects arising from exploratory analysis: Fit of word surprisal to LAN (top left) and N400 (top right) amplitudes; and fit of word entropy reduction to EPNP (bottom left) and PNP (bottom right) amplitudes, all as a function of the linguistic accuracy of the model that gave rise to the information measures. Dotted lines indicate the critical value of the χ^2-statistic for $\alpha = 0.05$. This is merely intended to provide an impression of the scale of χ^2 and should not be interpreted as indicating statistical significance, because significance has no meaning in an exploratory analysis.

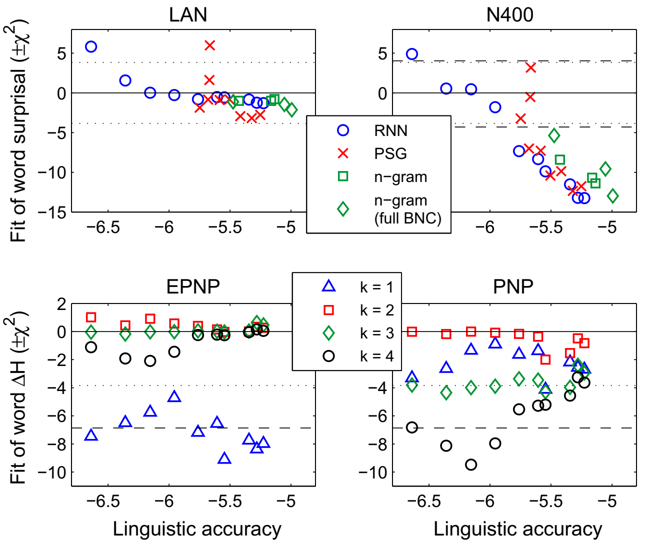

Fig. 2. Results from confirmatory analysis on the four potential effects: Fit of word surprisal to LAN (top left) and N400 (top right) amplitudes; and fit of word entropy reduction to EPNP (bottom left) and PNP (bottom right) amplitudes, all as a function of the linguistic accuracy of the model that gave rise to the information measures. Dotted lines indicate the critical value of the χ^2-statistic for $\alpha = 0.05$. Dashed lines are the levels beyond which effects are significant after applying Benjamini and Hochberg's (1995) multiple comparison correction procedure that controls the false-discovery rate at 5%.

analyses for content (i.e., open-class) and function (closed-class) words, constituting 53.2% and 46.8% of the data, respectively. A word's class was determined from its PoS tag, where nouns, verbs (including modal verbs), adjectives, and adverbs were considered content words, and all others were function words.

As can be seen in Fig. 4, there is no reliable N400 effect on function words. Nevertheless, the effect is generally weaker when only content words (as opposed to all words) are included. Most likely, this is because function words on average have lower surprisal and elicit a smaller N400 than content words. In other words, part of the effect over all words is due to the difference between content and function words.

3.3. Model comparison

Table 2 shows results of pairwise comparisons between the best models of each type, that is, those whose word surprisal estimates fit the N400 amplitude best (for a fair comparison with the RNN and PSG models, *n*-gram models trained on the full BNC were not included).

When looking at all words, the *n*-gram model's surprisal explains variance over and above each of the other two models whereas neither the RNN nor the PSG model significantly outperforms the *n*-gram. The RNN explains variance that the PSG does not account for, but the reverse is not the case. Taking only content words, results are similar except that the RNN now outperforms the *n*-gram model. Effects on function words are very weak in general and, consequently, no one model type accounts for variance over and above any other.

4. Discussion

If a word (or its part-of-speech) conveys more information, it takes longer to read the word. The first objective of the current study was to investigate whether ERP amplitude, too, depends on word and PoS information. Our expectation that the N400 would be related to word surprisal was indeed borne out. Other components and information measures, however, did not show any reliable correlation. Our second objective was to identify the model type whose information measures best predict the ERP data. Generally speaking, the *n*-gram and RNN models outperformed the PSG in this respect.

4.1. The N400 effect of word surprisal

Reading a word with higher surprisal value, under any of the three language model types, results in increased N400 amplitude. This finding confirms that the ERP component is sensitive to word predictability. Whereas previous studies (e.g., Dambacher et al., 2006; Kutas & Hillyard, 1984; Moreno, Federmeier, & Kutas, 2002; Wlotko & Federmeier, 2013) used subjective human ratings to quantify predictability, we operationalized (un)predictability as the information-theoretic concept of surprisal, as estimated by probabilistic language models that were trained on a large text corpus. Although word surprisal can be viewed as a more formal variant of cloze probability, it was not obvious in advance that the known effect of cloze probability on N400 size could be replicated by surprisal. As Smith and Levy (2011) demonstrated, systematic differences exist between cloze and corpus-based word probabilities, and cloze probabilities appear to predict word reading-times more accurately.

Across the full range of surprisal values, average N400 amplitudes differed by about 1 μV. Dambacher et al. (2006), too, found a difference of approximately 1 μV between content words with lowest and highest cloze probability. Experiments in which only sentence-final words are varied typically result in much larger effect sizes, with N400 amplitude varying by about 4 μV between high- and low-cloze (but not semantically anomalous) words (Kutas & Hillyard, 1984; Wlotko & Federmeier, 2013). Most likely, this is because effects are more pronounced on sentence-final words, or because cloze differences tend to be larger in handcrafted experimental sentences than in our (and Dambacher et al.'s) naturalistic materials.

All model types could account for the N400 effect as long as their linguistic accuracy was sufficient. Importantly, the strength of the surprisal effect grows nearly monotonically with linguistic accuracy, that is, models that more accurately embody the statistical patterns of English generate surprisal values that more accurately predict the N400. This finding validates our approach of extracting formal measures from corpus-based language models. Moreover, the relation between linguistic accuracy and amount of explained variance makes it very unlikely that the effect on the N400 is in fact due to a confounding variable rather than to surprisal per se. This is because such a confound would need to explain not only the effect of surprisal but also the effect of linguistic accuracy. The relation between N400 and word surprisal is further confirmed by the results of a recent fMRI study in which participants listened to spoken narratives (Willems, Frank, Nijhof, Hagoort, & Van den Bosch, 2014). Words with higher surprisal resulted in increased activation of the left temporal lobe, an area that has repeatedly been identified as an important source for the N400 (Service, Helenius, Maury, & Salmelin, 2007; Simos, Basile, & Papanicolaou, 1997; Van Petten & Luka, 2006).

N400 effects are usually investigated on content words only; Dambacher et al. (2006), too, excluded function words in their study of the relation between cloze probability and the N400. However, several studies have found that less predictable *function* words also result in increased N400 size (DeLong, Urbach, & Kutas, 2005; Martin et al., 2013; Wicha, Moreno, & Kutas, 2003). Separate analyses on content and function words revealed that, in our data, the effect is mostly (if not exclusively) present on content words. One reason why we failed to find a reliable N400 effect on function words might simply be that natural language (as captured in our

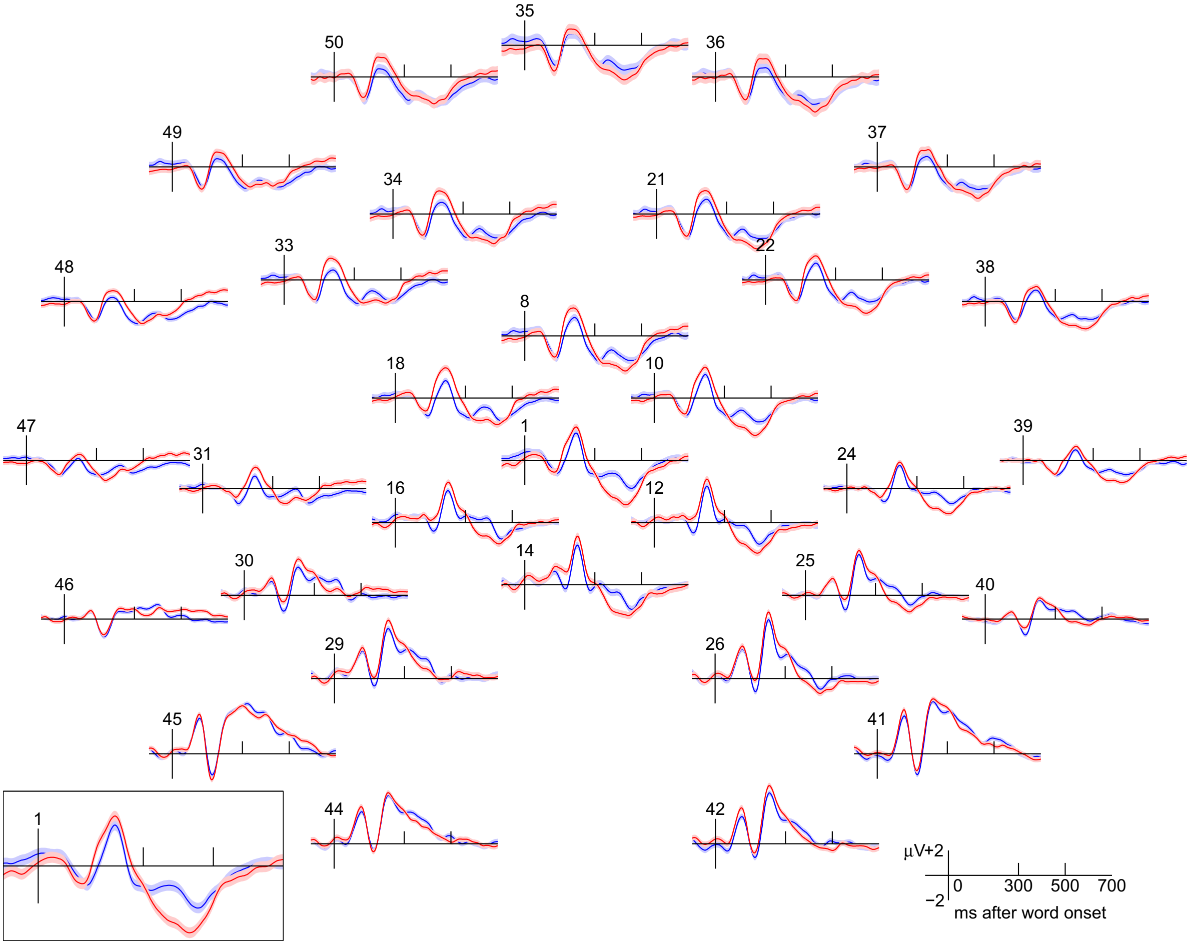

Fig. 3. Average electrode potentials, 0.50 Hz high-pass filtered and time-locked to word onset. Negative is plotted down. Numbers refer to electrode locations in montage M10 and correspond to those in Table 1. Red and blue lines are averages over the top and bottom tertile of word surprisal, respectively, and shaded areas represent 95% confidence intervals. Plots are baseline corrected such that the average potential over *both* tertiles equals zero in the 100 ms leading up to word onset (i.e., pre-onset differences between the two tertiles remain visible in the plots). The bottom left panel shows an enlarged plot for electrode 1.

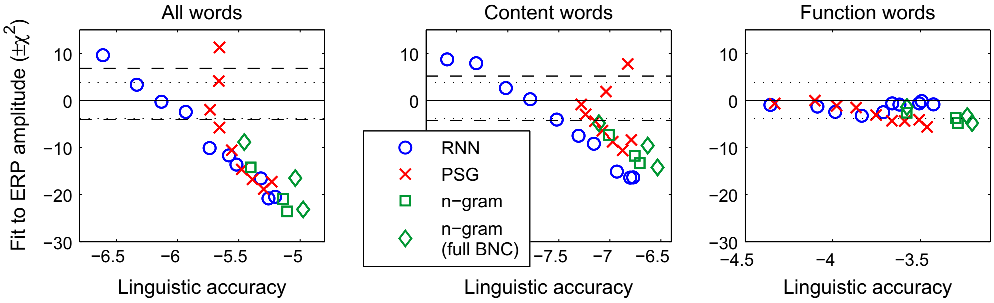

Fig. 4. Fit of word surprisal to N400 amplitude, on full data set (left), content words only (middle), or function words only (right), all as a function of the linguistic accuracy of the model that gave rise to the information measures. Note that linguistic accuracy depends on the data set because it is computed only over the words that take part in the analysis. Dotted lines indicate the critical value of the χ^2-statistic for $\alpha = 0.05$. Dashed lines are the levels beyond which effects are significant after applying Benjamini and Hochberg's (1995) multiple comparison correction procedure that controls the false-discovery rate at 5%.

sample of sentences) does not display much variance in function-word surprisal.

The question remains why word surprisal would be predictive of N400 size. Two functional interpretations of the N400 that have been proposed are that it reflects semantic integration (e.g., Hagoort, Baggio, & Willems, 2009; Kuperberg, 2007) or the retrieval of lexical information from memory (e.g., Brouwer, Fitz, & Hoeks, 2012; Kutas & Federmeier, 2000), with increased integration or retrieval difficulty resulting in a larger N400. We do not propose a third account but take the effect of surprisal to be subsumed by the memory-retrieval account: More predictable words can be pre-activated, thereby making it easier to retrieve their lexical information. In contrast, it is less clear why a more surprising word would be harder to semantically integrate into its sentence context, in particular when surprisal is estimated by language models that are only minimally (if at all) sensitive to semantics, as was the case

Table 2
Pairwise comparisons between word surprisal estimates by the best models of each type. Shown are the results of likelihood-ratio tests for the effect of one set of surprisal estimates (rows) over and above the other (columns). These tests are performed separately on the full data set, data on content words only, and data on function words only.

Words	Model	*n*-gram	RNN	PSG
All	*n*-gram		$\chi^2 = 6.01,\ p < .02$	$\chi^2 = 8.65,\ p < .01$
	RNN	$\chi^2 = 3.25,\ p < .08$		$\chi^2 = 4.44,\ p < .04$
	PSG	$\chi^2 = 1.80,\ p > .15$	$\chi^2 = 2.44,\ p > .1$	
Content	*n*-gram		$\chi^2 = 2.34,\ p > .1$	$\chi^2 = 3.68,\ p < .06$
	RNN	$\chi^2 = 6.98,\ p < .01$		$\chi^2 = 6.29,\ p < .02$
	PSG	$\chi^2 = 2.55,\ p > .1$	$\chi^2 = 0.52,\ p > .4$	
Function	*n*-gram		$\chi^2 = 1.97,\ p > .15$	$\chi^2 = 0.80,\ p > .3$
	RNN	$\chi^2 = 0.45,\ p > .5$		$\chi^2 = 0.22,\ p > .6$
	PSG	$\chi^2 = 1.65,\ p > .15$	$\chi^2 = 2.59,\ p > .1$	

here. The word probabilities estimated by our models arise from statistical word-order patterns, which depend much more on syntactic than on semantic factors.

4.2. Other ERP components and information measures

Gouvea, Phillips, Kazanina, and Poeppel (2010) argue that surprisal and entropy reduction, being 'one dimensional measures of syntactic processing cost' (p. 182), are unable to account for the variety in ERP components that can be extracted from the EEG signal. As we have seen, however, there are in fact many dimensions to these information-theoretic measures. Not only can each be estimated by many different probabilistic language models, we can also distinguish the dimensions of surprisal and entropy reduction, and of word and part-of-speech information. However, we did not find reliable ERP effects of entropy reduction, nor of the PoS-based measures. This null finding may be interesting in its own right, considering that all four information measures have been shown to account for word reading times. Frank (2013) attempted (and failed) to tease apart the individual reading-time contributions of word surprisal and entropy reduction and concluded that the two measures may not correspond to cognitively distinct processes. Instead, they would merely be alternative quantifications of one and the same cognitive factor. In that case, however, one would expect both of them to predict N400 amplitude. Our results suggest otherwise: Only word surprisal showed an effect, so this information measure appears to quantify neurally (and, most likely, cognitively) different processes than entropy reduction does.

Of course, we would have been able to draw stronger conclusions about the cognitive relevance of different information measures if they had accounted for different ERP components. Crucially, the absence of other effects is not due to problems with the EEG data (since an N400 effect was found) or the information measures (since these can explain reading times). This raises the question: Was there any reason to expect more than the N400 effect to begin with?

4.2.1. Left anterior negativities

It has been claimed that an ELAN effect occurs when the construction of a syntactic phrase structure fails (Friederici, Steinhauer, & Frisch, 1999; Gunter, Friederici, & Hahne, 1999; Neville, Nicol, Barss, Forster, & Garrett, 1991). More specifically, Lau, Stroud, Plesch, and Philips (2006) present evidence that an ELAN is elicited by the mismatch between the structural prediction based on the sentence so far and the syntactic category of the word currently being processed. This suggests that we may have found ELAN effects of PoS surprisal because this measure can be viewed as the extent to which a predicted syntactic category did not appear.

However, there are also several reasons why an ELAN effect was unlikely to arise. For one, it has been claimed that an ELAN only appears in cases of outright syntactic violations (Friederici, 2002; Friederici & Weissenborn, 2007), whereas all our experimental sentences are grammatically correct. Moreover, in a recent review of ERP studies on stimuli with local syntactic violations, Steinhauer and Drury (2012) concluded that an ELAN is more often absent than present in experiments that use visually presented sentences. They also argued that many of the studies that do find ELAN effects are methodologically flawed.

The LAN component is much less controversial than the ELAN. It appears to be elicited by a range of syntactic violations beyond the local phrase structure, such as subject-verb number disagreement and incorrect case marking (for a review, see Friederici & Weissenborn, 2007). However, LAN effects are not restricted to syntactic violations (Kaan & Swaab, 2003a) so, to the extent that syntactic difficulty is captured by word information, we could have observed a LAN effect in our data.

4.2.2. Late positivities

In a review of the literature, Van Petten and Luka (2012) write that most ERP studies that compare higher- and lower-cloze (but semantically congruent) words find not only the N400 but also an (E)PNP in response to the lower-cloze word. Hence, there was every reason for our word surprisal measure to predict the (E)PNP as well. In fact, results by Thornhill and Van Petten (2012) suggest that surprisal should be more predictive of the (E)PNP than of the N400: They found that presenting a low-cloze (i.e., high surprisal) synonym of a highly expected word elicits an (E)PNP but no N400.

Kaan and Swaab (2003b) found an anterior post-N400 positivity, much like the (E)PNP, in response to syntactic disambiguation. Be reminded from the Introduction that entropy reduction can be viewed as the amount of ambiguity resolved by a word or PoS. Therefore, entropy reduction might predict the (E)PNP. Indeed, our exploratory analysis did reveal a potential (E)PNP effect of word entropy reduction, which closely followed findings on reading time in that the effect grew stronger with higher linguistic accuracy and larger lookahead distance. Somewhat disappointingly, no such effect remained in the confirmatory analysis. Although this striking difference between the two data sets may well be a statistical fluke, it raises the question if there was any relevant difference between the subject groups of the two analyses. There were no large differences in either mean age or gender (Exploratory: 29.5 years, 6 females; Confirmatory: 26.4 years, 4 females) but the groups might have differed in other properties. In any case, the possible effect of entropy reduction on (E)PNP deserves further study.

The P600, which is a more posterior component than the (E)PNP, is well known to occur when there is a syntactic garden path (e.g., Kaan & Swaab, 2003b; Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994). This has given rise to claims that it reflects a process of syntactic reanalysis that takes place when an initial sentence interpretation turns out to be incorrect (Friederici, 2002; Osterhout et al., 1994). A garden-path effect is necessarily triggered by the appearance of a word with an unexpected syntactic category. As such, syntactic reanalysis should co-occur with high surprisal and, indeed, surprisal has been shown to account for particular garden path phenomena (Brouwer, Fitz, & Hoeks, 2010; Hale, 2001). Levy (2008) proves that surprisal equals the extent to which the current input forces reallocation of the probability assigned to possible sentence interpretations. In other words, surprisal is large if a highly probable interpretation turns out to be incorrect (i.e., a garden path is encountered) and the associated probability must be reallocated to other (previously unlikely) interpretations. If the P600 indeed reflects syntactic reanalysis, we could therefore have seen surprisal effects on the P600. Even an entropy-reduction effect could not have been excluded in advance, considering that Hale (2003) and Linzen and Jaeger (2014) demonstrate that some garden paths can be viewed as effects of entropy reduction rather then surprisal. However, the P600 has also been found in cases that do not involve increased syntactic processing difficulty (e.g., Hoeks, Stowe, & Doedens, 2004; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Regel, Gunter, & Friederici, 2011; Van Berkum, Koornneef, Otten, & Nieuwland, 2007). This led to alternative interpretations of the P600 effect (e.g., Brouwer et al., 2012; Kuperberg, 2007) in which syntactic processing plays no central role and there is no reason to expect any effect of information quantities (at least, not as captured by our language models).

4.3. Implications for models of sentence comprehension

Cloze probabilities depend not only on participants' knowledge of language but also on non-linguistic factors, such as world knowledge and metacognitive strategies. Our model-derived probabilities are very different in this respect, because they are solely based on the statistical language patterns extracted from the training corpus. Consequently, the use of computational models (as opposed to cloze probabilities) allows us to isolate purely linguistic effects on the EEG signal. More importantly, evaluating and comparing the predictions by structurally different models against the same set of experimental data provides insight into the cognitively most plausible sentence comprehension processes.

Model comparisons revealed significant differences between model types with respect to the N400 effect. In particular, the *n*-gram and RNN model accounted for variance in N400 size over and above the PSG whereas the reverse was not the case. In short, the more parsimonious models, which do not rely on assumptions specific to language, outperform the hierarchical grammar based system. This mirrors results from reading time studies (Frank & Bod, 2011; Frank & Thompson, 2012; but see Fossum & Levy, 2012), suggesting that the assumptions underlying the PSG model are not efficacious for generating expectations about the upcoming word. Such a conclusion is consistent with claims that a non-hierarchical, RNN-like architecture forms a more plausible cognitive model of language processing than systems that are based on hierarchical syntactic structure (e.g., Bybee & McClelland, 2005; Christiansen & MacDonald, 2009; Frank, Bod, & Christiansen, 2012).

Likewise, it is noticeable that there was no effect on ERP components that are traditionally considered to reflect syntactic processing effort. Although no strong conclusions should be drawn from such a null effect, an effect of PSG-based surprisal (perhaps particularly for parts-of-speech) would be expected if the construction of a syntactic structure is fundamental to sentence comprehension.

5. Conclusion

This work connected information-theoretical notions to their neural implementations, revealing a strong relation between the surprisal of a word and the amplitude of the N400 component in response to reading that word. Evidently, information quantities derived from statistical language models can be used to make sense of EEG data from large-scale, non-factorial studies that use naturally occurring sentences as stimuli. This offers a novel technique for setting-up and analyzing EEG studies, one that does not rely on the careful construction of stimuli and manipulation of factors.

Any probabilistic language model can be used to estimate word information values, allowing for a very flexible approach to model evaluation and comparison which can be instrumental in uncovering the representations and processes that underlie human sentence processing. The three types of models we used here are relatively simple; more sophisticated systems are likely to be better capable at simulating cognitive processes. Future modeling efforts may therefore result in more appropriate information estimates to evaluate against EEG data, possibly revealing novel correspondences between information values and ERP responses. To facilitate such future endeavors, we make our data available as online supplementary materials to the research community. We hope and expect that formal modeling can help shed light on the oftentimes contradictory-seeming ERP findings.

Acknowledgments

We would like to thank Elisabet Service and an anonymous reviewer for their helpful comments on an earlier version of this paper.

The current article is an extended and improved version of a paper presented at the 51st Annual Meeting of the Association for Computational Linguistics (Frank, Otten, Galli, & Vigliocco, 2013). The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007–2013) under a Marie Curie Intra-European Fellowship (Grant No. 253803) and a Career Integration Grant (Grant No. 334028), both awarded to the first author. The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility, and associated support services, in the completion of this work.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bandl.2014.10.006.

References

Barr, D. J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

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.

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*, 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes, 26*, 301–349.

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review, 22*, 381–410.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13*, 359–394.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning, 59*, 126–161.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effect on event-related potentials during reading. *Brain Research, 1084*, 89–103.

De Groot, A. D. (2014). The meaning of significance for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, Han L.J. van der Maas]. *Acta Psychologica, 148*, 188–194.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*, 1117–1121.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*, 193–210.

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

Fernandez Monsalve, I., 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). Avignon, France: Association for Computational Linguistics.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–69). Montréal, Canada: Association for Computational Linguistics.

Frank, S. L (2010). Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 81–89). Uppsala, Sweden: Association for Computational Linguistics.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive processing load in sentence comprehension. *Topics in Cognitive Science, 5*, 475–494.

Frank, S. L. (2014). Modelling reading times in bilingual sentence comprehension. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1860–1861). Austin, TX: Cognitive Science Society.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*, 829–834.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences, 279*, 4522–4531.

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

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 878–883). Sofia, Bulgaria: Association for Computational Linguistics.

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.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*, 78–84.

Friederici, A. D., Steinhauer, K., & Frisch, S. (1999). Lexical integration: Sequential effects of syntactic and semantic information. *Memory & Cognition, 27*, 438–453.

Friederici, A. D., & Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax-semantics interface. *Brain Research, 1146*, 50–58.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes, 25*, 149–188.

Gunter, T. C., Friederici, A. D., & Hahne, A. (1999). Brain responses during sentence reading: Visual input affects central processes. *NeuroReport, 10*, 3175–3178.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 819–836). Cambridge, MA: MIT Press.

Hale, J. T. (2001). A probabilistic early parser as a psycholinguistic model. *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hale, J. T. (2003). The information conveyed by words. *Journal of Psycholinguistic Research, 32*, 101–123.

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30*, 643–672.

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science, 35*, 399–443.

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*, 59–73.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

John, J. A., & Draper, N. R. (1980). An alternative family of transformations. *Applied Statistics, 29*, 190–197.

Kaan, E., & Swaab, T. Y. (2003a). Electrophysiological evidence for serial sentence processing: A comparison between non-preferred and ungrammatical continuations. *Cognitive Brain Research, 17*, 621–635.

Kaan, E., & Swaab, T. Y. (2003b). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*, 98–110.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430). Sapporo, Japan: Association for Computational Linguistics.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23–49.

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language, 100*, 223–237.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*, 463–470.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163.

Lau, E., Stroud, C., Plesch, S., & Philips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language, 98*, 74–88.

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

Linzen, T., & Jaeger, F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Baltimore, MD: Association for Computational Linguistics.

Martin, C. D., Thierry, G., Kuipers, J., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*, 574–588.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 196–206). Uppsala, Sweden: Association for Computational Linguistics.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language, 80*, 188–207.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience, 3*, 151–165.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research, 1118*, 155–167.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*, 785–806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 786–803.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and Latent Semantic Analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46). Canberra, Australia.

Regel, S., Gunter, T. C., & Friederici, A. D. (2011). Isn't it ironic? An electrophysiological exploration of figurative language processing. *Journal of Cognitive Neuroscience, 23*, 277–293.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics, 27*, 249–276.

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* (pp. 324–333). Association for Computational Linguistics.

Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank project* (Tech. Rep.). Philadelphia, PA: University of Pennsylvania.

Service, E., Helenius, P., Maury, S., & Salmelin, R. (2007). Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience, 19*, 1193–1205.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. 623–656.

Simos, P. G., Basile, L. F. H., & Papanicolaou, A. C. (1997). Source localization of the N400 response in a sentence-reading paradigm using evoked magnetic fields and magnetic resonance imaging. *Brain Research, 762*, 29–39.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1637–1642). Austin, TX: Cognitive Science Society.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language, 120*, 135–162.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 901–904). Denver, Colorado.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*, 382–392.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research, 1146*, 158–171.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97*, 279–293.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex, 39*, 483–508.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2014). Prediction during natural language comprehension (submitted for publication).

Wlotko, E. W., & Federmeier, K. D. (2013). Two sides of meaning: The scalp-recorded N400 reflects distinct contributions from the cerebral hemispheres. *Frontiers in Psychology, 4*(181).

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1189–1198). Uppsala, Sweden: Association for Computational Linguistics.