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/

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