Anticipation in language comprehension: prediction and priming show distinct patterns of brain activity

Stefan L. Frank[1] and Roel M. Willems[1,2,3]

[1]Centre for Language Studies, Radboud University Nijmegen
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen
[3]Max Planck Institute for Psycholinguistics

Abstract

Word anticipation during language comprehension can, in principle, proceed via probabilistic prediction or via semantic priming. We investigate if these two pathways are neurally distinct by reanalyzing EEG and fMRI data from studies in which participants comprehend naturalistic stimuli. Each content word's predictability given previous words is quantified by a probabilistic language model, and semantic relatedness to previous words is quantified by a distributional semantics model. Brain activity time-locked to each word is regressed on the two model-derived measures. Results show that prediction and semantic priming have near identical N400 effects but are dissociated in the fMRI data, with word prediction related to activity in, among others, the visual word-form area and priming related to activity in areas associated with the semantic network. This indicates that prediction and priming both occur during natural language comprehension and are cognitively distinct processes.

1 Introduction

1.1 Pathways to word anticipation

Anticipatory processing is known to play an important role during language comprehension (for recent reviews, see Huettig, 2015, and Kuperberg & Jaeger, in press). Word anticipation is usually thought of as the result of (probabilistic) predictions that are based on contextual information and knowledge of the language and the world. To rehash a famous example, Altmann and Kamide (1999) had participants listen to sentences like 'The boy will eat the –' while viewing an image containing several objects, among which a boy and a cake but no other edible object. Immediate looks at the cake revealed prediction of the upcoming word 'cake'. Such prediction requires knowledge of the language (e.g., the SVO structure of English) and the world (cakes are edible) as well as the use of information from the linguistic and non-linguistic context

(the spoken sentence and visually presented objects, respectively).

In the current study, we investigate the additional role of an alternative pathway to word anticipation, one that relies not on the structure of language and the world but on semantic relations between words: semantic priming. By *anticipation*, we will refer to any process that increases the ease with which an upcoming word can be accessed or processed. By *prediction*, we mean any anticipatory process that relies on knowledge of the structure of the language and the world, although we limit ourselves to linguistic context and linguistic knowledge in the current work. By *priming*, we mean any anticipatory process that relies on semantic relations between the upcoming word and earlier encountered words.[1]

Because prediction is knowledge-driven, it is commonly viewed as a top-down cognitive process. In contrast, priming is traditionally conceptualized as a process of bottom-up activation spreading in a semantic or associative network (Collins & Loftus, 1975). Another way to think about the difference between prediction and priming (as we define them) is in terms of syntagmatic and paradigmatic relations in language. The words 'boy', 'eat', and 'cake' are syntagmatically related: They need to occur in this order because English has SVO structure and 'eat' is a verb for which 'boy' is an appropriate subject and 'cake' an appropriate object. The three words are not paradigmatically related because they cannot take each other's place without radically changing the meaning of the sentence ('the cake will eat the boy') or making it ungrammatical ('The eat will boy the cake'). However, the word 'cake' is paradigmatically related to, for example, 'pie' because 'pie' can take the place of 'cake' in most contexts.

It should be clear that prediction makes use of syntagmatic relations ('cake' is likely to occur after 'the boy will eat the –') whereas priming is based on paradigmatic relations ('cake' can prime 'pie' because of their semantic relatedness). Hence, prediction and priming are conceptually distinct cognitive processes applied to different representations. The objective of the current study is to reveal whether prediction and priming are also neurally distinct, which would show that the cognitive system indeed makes use of these two processes to anticipate words during language comprehension.

1.2 Relation to earlier work

At first glance, predictability and semantic relatedness may appear to correspond to, respectively, cloze probability and semantic congruency, which have been the focus of many EEG and fMRI studies on sentence comprehension. However, the relation between these two pairs of variables is more complex.

To obtain cloze probabilities, participants in a cloze task name what they consider to be most likely upcoming word of an incomplete sentence. If semantic priming increases word anticipation during sentence comprehension, they will be affected by the semantic relatedness between words. Hence, cloze probabilities may depend not only on linguistic and world knowledge but also one semantic relatedness, whereas our notion of predictability explicitly excludes relatedness.[2] In short, cloze probability equals word predictability plus

[1] Kuperberg and Jaeger (in press) make the same distinction between predictive processes and priming.

[2] There is one caveat: If semantic priming affects language production, this will increase the co-occurrence of semantically

the effect of semantic priming.

Semantic congruency depends on the relation between sentence meaning and what is possible in the real (or narrative) world. Hence, world knowledge makes more congruent words more predictable, independently from their syntagmatic or paradigmatic relatedness with the sentence context. In short, semantic congruency equals word predictability minus the effect of linguistic knowledge.

Earlier studies that attempt to dissociate prediction and priming differ in several important respects from the current work. Lau, Holcomb, and Kuperberg (2013) use a semantic decision task with prime-target word pairs, where predictability of a target's semantic class is varied by manipulating the proportion of related-prime fillers in an experimental block. They take prediction of a word or semantic feature to come down to commitment in working memory whereas priming increases activation in long-term memory. We refrain from assumptions regarding the memory structures involved but take prediction and priming to differ in the type of knowledge involved (syntagmatic and paradigmatic, respectively). In contrast, Lau et al. (2013) make clear that they take prediction and priming to rely on the same knowledge.

Following Lau et al. (2013) in assuming that word prediction comes down to commitment, Brothers, Swaab, and Traxler (2015) aim to 'dissociate the effects of *specific* lexical pre-activation, from other sources of contextual support (e.g., semantic association or discourse plausibility)' (p. 136). In our terminology, 'semantic association' would be a source of priming whereas 'discourse plausibility' can lead to increased word anticipation due to prediction. Hence, Brothers et al. (2015) conflate these two processes and contrast them to an all-or-nothing 'lexical pre-activation', which is operationalized by instructing participants to explicitly predict the final word of a short discourse and indicate whether their prediction matched the actual final word.

Unlike Brothers et al. (2015) and Lau et al. (2013), we do not assume that prediction involves any sort of commitment to a lexical item or semantic feature. Rather, both prediction and priming change the extent to which individual words are anticipated. If one word (or a set of words sharing a semantic feature) is much more strongly anticipated than all others, this may be viewed as a commitment to that word or feature but this is not qualitatively different from just strong anticipation.

In an ERP study, Metusalem et al. (2012) had participants read short narratives where the critical word in the final sentence was either highly expected or semantically anomalous, and semantically anomalous words were either related or unrelated to the event described in the text. The N400 response to anomalous words was smaller in the event-related condition, suggesting that even highly unpredictable (i.e., zero cloze probability) words are to some extent anticipated if they are associated to the earlier discourse.

Our study's methodology differs from Brothers et al.'s (2015), Lau et al.'s (2013), and Metusalem et al.'s (2012) in two ways. First, we make use of EEG/fMRI recordings of participants' brain activity during normal reading/listening comprehension of naturally occurring sentences or texts, rather than items constructed for

related words in language, that is, they become syntagmatically related. In this indirect way, semantic relatedness can come to affect predictability via knowledge of language structure.

the sake of the experiment. Consequently, the materials do not contain any semantic anomalies. Second, the EEG or fMRI signal at each content word is compared to measures of the word's 'predictability' and 'primeability', derived from computational models that quantify the syntagmatic and paradigmatic relatedness between a content word and previous words in the stimuli. In this manner, we are able to tease apart effects of prediction and priming (as operationalized by the computational models) without explicitly manipulating (cloze) probability or semantic relatedness.

1.3 Models of syntagmatic and paradigmatic relations in language

In the field of Computational Linguistics, a *language model* is by definition any probability model that assigns probabilities to sentences or, equivalently, assigns a conditional probability distribution $P(w_t|w_1, \ldots, w_{t-1})$ over the potentially upcoming words w_t given the sequence of words so far $w_1, \ldots, w_{t-1}$.[3] Word probability can easily be transformed to *word surprisal*, defined as $-\log P(w_t|w_1, \ldots, w_{t-1})$, which has been argued to form a cognitively relevant measure of processing load when encountering word w_t in sentence context (Hale, 2001; Levy, 2008). Indeed, it has repeatedly been shown that surprisal predicts word-reading time (Frank & Thompson, 2012; Monsalve, Frank, & Vigliocco, 2012; Smith & Levy, 2013). Surprisal effects have also be found in brain imaging data: Higher surprisal value results in a stronger N400 ERP component (Frank, Otten, Galli, & Vigliocco, 2015) as well as its MEG equivalent (Parviz, Johnson, Johnson, & Brock, 2011; Wehbe, Vaswani, Knight, & Mitchell, 2014), and stronger BOLD response in anterior temporal cortex, inferior frontal gyrus, and the visual word-form area (Hale, Lutz, Luh, & Brennan, 2015; Willems, Frank, Nijhof, Hagoort, & Van den Bosch, in press).

In the current study, we use word surprisal as a formalization of the extent to which the word is (or, rather, can be) anticipated using top-down prediction. In contrast, to quantify potential anticipation due to bottom-up semantic priming, we make use of a distributional lexical semantic model. Such a model captures paradigmatic relations between words by keeping track of word co-occurrence in a large text corpus. Many distributional semantic models have been proposed, the best known being Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). More recent approaches include Bayesian models (Griffiths, Steyvers, & Tenenbaum, 2007) and neural networks (Mikolov, Chen, Corrado, & Dean, 2013). In each case, words are represented by high-dimensional vectors and distances between vectors quantify the semantic distances between words. These distances have been shown to be predictive of priming effects in naming (Jones, Kintsch, & Mewhort, 2006) and lexical decision (Günther, Dudschig, & Kaup, in press) experiments.

In the context of sentence or text comprehension, computationally quantified semantic distance has been studied much less than surprisal. Pynte, New, and Kennedy (2008) found that larger LSA distance between the current and previous content word(s) results in longer word-reading time. Although they do not factor out any measure of word probability or surprisal, they do argue that the semantic distance effect cannot

[3]The probability of w_t could also depend on non-linguistic context, such as the current visual scene, but few language models take non-linguistic information into account.

be reduced to an effect of word predictability. To the best of our knowledge, there have not been any neuroimaging studies that look at effects of semantic relatedness (as quantified by computational models) during the comprehension of naturalistic stimuli. However, several reading studies have looked at effects of semantic relatedness using a factorial experimental design and hand-crafted sentence stimuli and found priming effects on the N400 (Camblin, Gordon, & Swaab, 2007; Stafura & Perfetti, 2014; Van Petten, 1993). Analysing MEG data on sentence-final words occurring in high and low constraining context pairs, Parviz et al. (2011) showed that N400 strength correlates positively with word surprisal as well as LSA-based semantic distance to the sentence's previous words. Others have found neural correlates of the semantic vectors themselves (rather than distances between them), both in single word comprehension (T. Mitchell et al., 2008) and in narrative reading (Wehbe, Murphy, et al., 2014; Wehbe, Vaswani, et al., 2014).

As an alternative to capturing either syntagmatic or paradigmatic relations of language, a few models combine both into a single system. For example, Jones and Mewhort's (2007) BEAGLE model constructs word vectors that capture not only semantic relatedness but also word-order information. The reversed approach (including paradigmatic structure in a probabilistic next-word prediction model) is more common. Indeed, semantic vector representations can be incorporated into a language model to improve its surprisal estimates. J. Mitchell and Lapata (2009) developed a model that explicitly uses semantic distance values to adjust word-probability estimates and J. Mitchell, Lapata, Demberg, and Keller (2010) show that this not only improves the language model but also provides surprisal values that more accurately predict reading times. Recurrent Neural Network (RNN) language models, trained to perform next-word prediction, will automatically develop vector representations of words in their input connection weights, thereby capturing the words' semantic relatedness (Brakel & Frank, 2009; Mesnil, He, Denk, & Bengio, 2013; Mikolov et al., 2013). Surprisal values by RNNs are therefore based on both syntagmatic and paradigmatic relations.

The Mitchell et al. and RNN models implement an alternative to the two-pathway view of word anticipation: Semantic relatedness between words only increases anticipation via the prediction system. That is, the cognitive systems considers the occurrence of semantically related words more likely and therefore predicts their occurrence, which is simulated in a language model by assigning a lower surprisal to semantically related words. In that case, we would expect that word surprisal (under a language model that does not incorporate paradigmatic relations) and semantic relatedness have identical effects on brain activity. Conversely, if surprisal and relatedness have dissociable effects, this would support the two-pathway view of anticipation. A third possibility, of course, is that model-derived semantic relatedness measures have no measurable effect on brain activity during language comprehension.

1.4 The current study

In what follows, we will briefly describe the previously published EEG and fMRI data sets in which surprisal and relatedness effects will be identified (Section 2.1) after which we explain the two models that estimated these formal measures (Section 2.2). A large-scale regression analysis is then applied to identify unique

effects of surprisal and relatedness in the neuroimaging data, over and above a number of covariates. If prediction and priming are neurally distinct processes, we should find that the two measures differ in the timing or distribution of their ERP effects or in the associated brain areas as measured with fMRI. The results (Section 3) show near identical effects of of prediction and priming on EEG, at least as far as the N400 is concerned, but a clear difference in the fMRI data but This strongly suggests that word anticipation involves both the top-down prediction and bottom-up priming pathways.

2 Method

2.1 Neuroimaging data

We reanalyzed data from two publications on neuroimaging during language comprehension, one an EEG study (Frank et al., 2015) and the other applying fMRI (Willems et al., in press). We only briefly discuss the stimuli materials and data collection here. Full details can be found in the original papers.

In the EEG study, 24 native speakers of English read 205 individual sentences that were sampled from English narratives, presented centrally one word at a time (RSVP method) with a word-length-dependent SOA of at least 627 ms. All sentences contained at least two content words (see Frank, Monsalve, Thompson, & Vigliocco, 2013, for the sentence selection constraints). Comprehension was tested by means of yes/no questions that appeared after approximately 50% of the sentences. EEG was recorded on 32 channels at 500 Hz (downsampled to 250 Hz and band-pass filtered between 0.5 and 25 Hz) and epoched into trials from -100 ms to $+700$ ms relative to word onset. Trials with artifacts were identified visually and removed.

In the fMRI study, 24 native speakers of Dutch listened to three short excerpts from Dutch narrative audiobooks for a total of 19:27 minutes. The reversed audio files were presented as a baseline condition. There was no explicit task, but participants were tested post-hoc for their memory and comprehension of the narratives. Images of Blood-Oxygenation Level Dependent (BOLD) changes were acquired on a 3T Siemens scanner with a T2*-weighted 3D EPI sequence (Poser, Koopmans, Witzel, Wald, & Barth, 2010; TR: 880 ms, TE: 28 ms, flip angle: 14 degrees, voxel size: $3.5 \times 3.5 \times 3.5$ mm, 36 slices). Preprocessing involved motion correction (spatial realignment), spatial normalization to MNI space, and spatial smoothing (8 mm FWHM), using SPM8 (http://www.fil.ion.ucl.ac.uk/spm).

2.2 Quantifying word anticipation

As explained in detail below, each content word from the Dutch and English stimuli was characterized by two measures: suprisal and semantic distance. Surprisal quantifies the extent to which the word's occurrence is unexpected given the previous words and knowledge of syntagmatic relations in the language. Semantic distance quantifies the extent to which the current and previous content words tend to occur in different contexts, which requires knowledge of paradigmatic relations in the language.

Both the surprisal and semantic distance model were trained on the first slice of Corpora from the Web (Schäfer, 2015), comprising individual sentences from web sources. The Dutch NLCOW14 corpus comprises 37.0 million sentences with 683.6 million word tokens of 4.95 million types. The English ENCOW14 corpus comprises 28.9 million sentences with 644.5 million word tokens of 2.81 million types. Words include punctuation, numbers, and other non-verbal symbols; and word-type count is case-insensitive. The much larger number of word types in Dutch compared to English is mostly due the fact that noun-noun compounds are written as single words in Dutch.

2.2.1 Surprisal

We made use of an n-gram language model for generating word-surprisal values.[4] Such models are 'myopic' in that a word's probability estimate depends only on the previous $n-1$ words: The full conditional probability $P(w_t|w_1, \ldots, w_{t-1})$ is simplified to $P(w_t|w_{t-n+1}, \ldots, w_{t-1})$. Although more sophisticated language models can handle longer dependencies between words, we opted for the simple n-gram model because they are easy to train on very large data sets and less language- or theory-dependent than, for example, phrase-structure grammars. An alternative would be to use RNNs, which are able to capture long-distance dependencies (at least, in principle) and have been shown to outperform n-gram models (Mikolov, Karafiát, Burget, Černocký, & Khudanpur, 2010). However, as explained in the Section 1.3, they do so by developing vector representations of words that can capture semantic relations. As our current objective is to distinguish between effects of surprisal and semantic relatedness, the model that estimates surprisal values should not have access to any semantic information of the sort that is captured by semantic vector representations.

2.2.2 Semantic distance

Semantic vector representations of words were generated by Mikolov et al.'s (2013) skipgram model.[5] As illustrated in Figure 1, this is a two-layer feedforward neural network that receives as input individual words tokens from the training corpus and learns to produce as output the previous and upcoming five words (or less, respecting sentence boundaries) in the training sentence. Error is backpropagated through the network to update the connection weights, so that the weight vectors from two input nodes become more similar if the corresponding two words often occur in similar contexts, that is, if they are paradigmatically related. In this way, input weight vectors become semantic representations of words. Note that words that tend to co-occur (e.g., 'eat the cake') do not usually occur in similar contexts (i.e., 'the' follows 'eat' but precedes 'cake') so do not receive similar vector representations. Rather, their syntagmatic relatedness is captured by the n-gram surprisal model.

[4] This model was obtained using SRILM (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen & Goodman, 1999) and $n = 5$.

[5] Model settings were: initial learning rate 0.025; 5 negative samples; words with frequency above 10^{-3} were downsampled; 5 iterations through training data

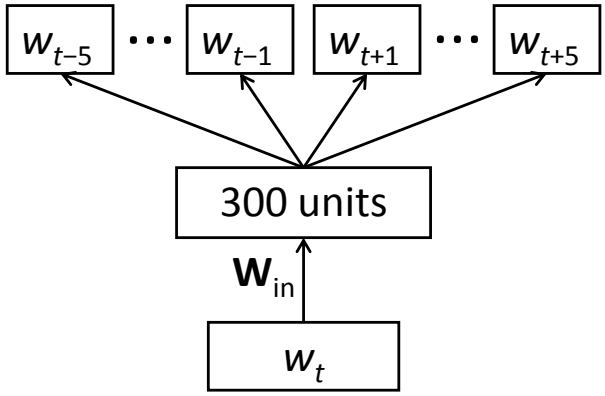

Figure 1: Architecture of the Mikolov et al. (2013) neural network model for obtaining word vector representations. The input layer and each of the 10 output blocks contain one unit for each word type in the training corpus. Each column of the input weight matrix $\mathbf{W}_{\text{in}}$ is the 300-dimensional vector that represents the corresponding word.

The strength of the semantic relation between the current word and previous content words is defined as the cosine of the angle between the current word vector and the sum of previous content word vectors. Two different inclusion criteria are applied when summing over previous word vectors. Under the 'narrow' criterion, only content words within the four words immediately preceding the current word are included. The 'wide' criterion allows for more words to be included in the vector sum. For the narrative texts of fMRI study, it includes the vectors of previous three content words (or fewer, for words at the very beginning of a text). For the individual sentences from EEG study, the wide criterion includes all previous content words in the sentence. Irrespective of the criterion, an item is discarded from analysis if it is the first content word of the text (fMRI study) or sentence (EEG study).

Under the narrow inclusion criterion, the semantic relatedness measure can only make use of context that is also available to the 5-gram surprisal measure (see Section 2.2.1) so it is ideal for direct comparison between surprisal and semantic relatedness effects. However, as the narrow window may include only one content word the previous context's semantics may not be accurately represented. Therefore, priming effects may be easier to discover using the wider window. Not surprisingly, the two measures are strongly correlated, both for the Dutch ($r = .68$) and the English materials ($r = .80$).

We expect higher cognitive load, as observed in a stronger BOLD or N400 signal, for words with higher surprisal or weaker semantic relatedness. To obtain effects in the same direction for both measures, we use the *negative* of relatedness as our measure of semantic *distance*. Surprisal and distance of content words in the Dutch materials did not correlate (narrow criterion: $r = .02$; wide criterion: $r = .03$) but there was a weak positive correlation in the English stimuli (narrow criterion: $r = .20$; wide criterion: $r = .27$).

3 Results

Both the EEG and fMRI data sets were analysed by regressing brain activity measures (electrode potential or BOLD response) on surprisal and semantic distance, but analysis details differed because of the type of data. Note that the surprisal and distance measures are included in the regression analysis together, so the effect of one measure is over and above what is already explained by the other.

3.1 EEG

As discussed in the Introduction, earlier EEG studies looking at semantic priming in sentence comprehension found effects on the N400. For this reason, we look only at the seven most central electrodes, where Frank et al. (2015) already found an N400 effect of surprisal in this EEG data set. The objectives here are to ascertain if semantic distance, too, affects the N400 and, if so, to compare its timing, distribution, and effect size to that of the surprisal-elicited N400 wave.

The analysis roughly followed the rERP method recently proposed by Smith and Kutas (2015) where the set of electrode potentials at each sample point, collected over word token and subjects, is regressed on the relevant predictor. The statistics of interest are then the regression coefficients of surprisal and semantic distance. Following Frank et al. (2015), the covariates included were: position of sentence in the experiment session, position of word in the sentence, word length, word frequency (in the training corpus, log transformed), and EEG baseline (average electrode potential in the 100 ms leading up to word onset). All independent variables were standardized. The mixed-effects regression model included by-subject and by-word random intercepts and by-subject random slopes of surprisal and semantic distance.

Figures 2 and 3 display the time course of the coefficients b from the regression analysis (expressed in μV per standard deviation increase in the predictor) for the effects of surprisal and semantic distance on electrode potential, time-locked to word onset. The two figures differ in the semantic distance inclusion criterion: 'narrow' for Figure 2 and 'wide' for Figure 3. In both cases, the effects of surprisal and semantic distance have nearly identical onset and offset, and even the effect sizes are very similar, that is, a unit increase in surprisal leads to the same change in electrode potential as a unit increase in one semantic distance, when both measures are expressed in standard deviations of the distribution over the stimuli materials. The difference between the effects of the 'narrow' and 'wide' semantic distance measures is minimal, although effects appear to be slightly weaker and more left lateralized for the 'narrow' compared to the 'wide' measure.

3.2 fMRI

The analysis of fMRI was identical to that in Willems et al. (in press) except semantic relatedness was included as a predictor in addition to word surprisal and frequency (in the training corpus, log transformed) and that function words, for which no semantic distance values are computed, were modelled as events of no interest. In contrast to the EEG study above, we had no clear prior expectations about where any effect of

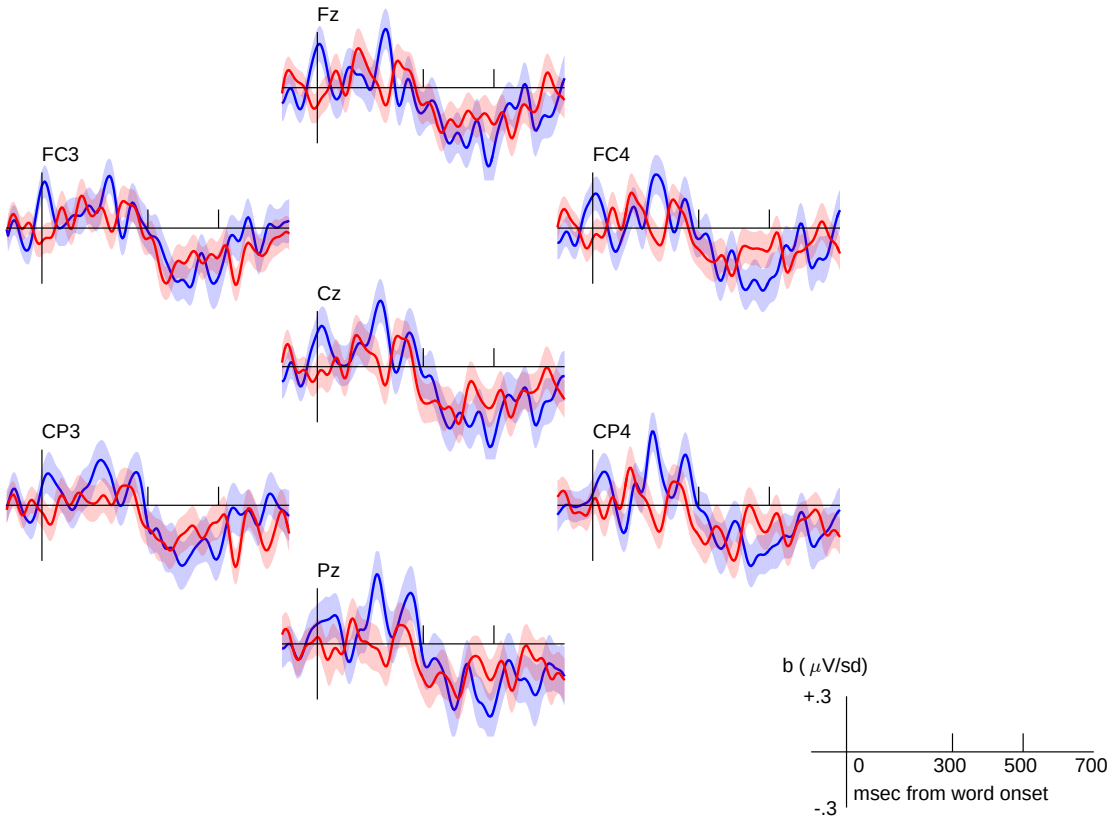

Figure 2: Regression coefficients (b) for the effects of surprisal (blue) and 'narrow criterion' semantic distance (red) on electrode potential, in each 4ms sample from −100ms to +700ms relative to word onset. Shaded areas indicate standard error.

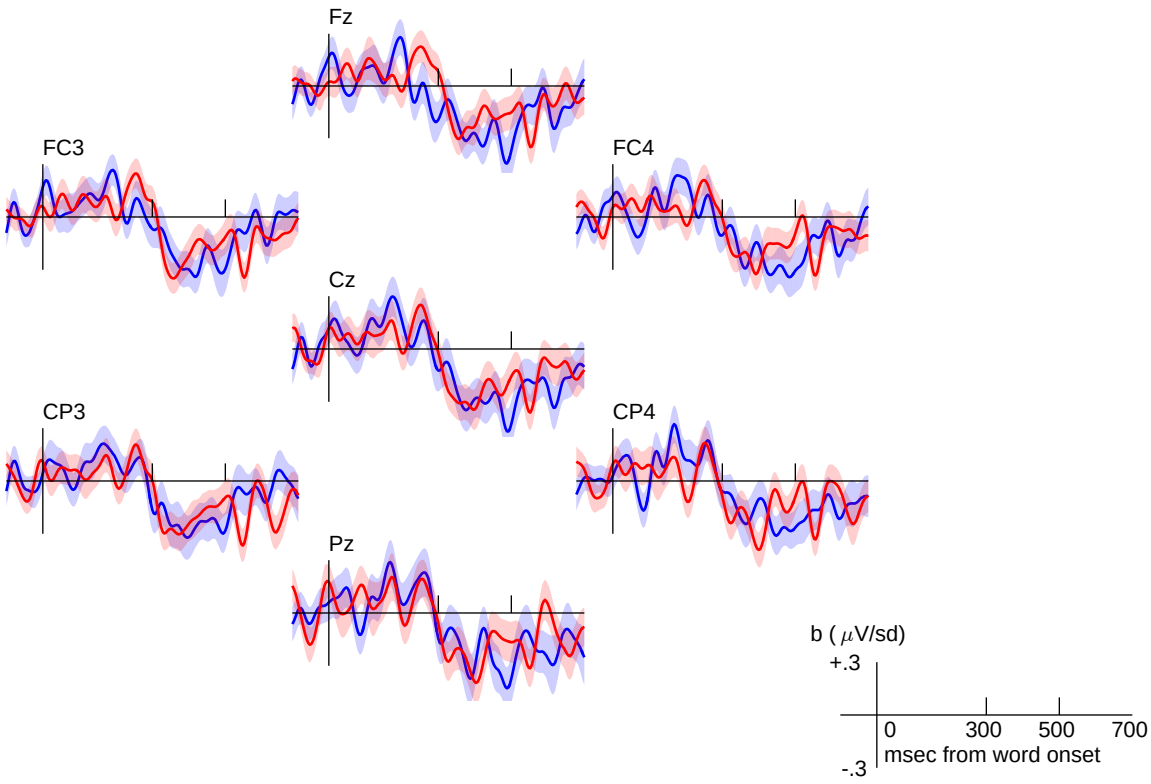

Figure 3: Regression coefficients (b) for the effects of surprisal (blue) and 'wide criterion' semantic distance (red) on electrode potential, in each 4ms sample from −100ms to +700ms relative to word onset. Shaded areas indicate standard error.

Table 1: Brain areas that become significantly more active in response to word surprisal. The table displays a description of the region, coordinates in stereotaxic MNI space, the extent of the activation cluster, and the t-value of the reported voxels. Large clusters are represented by two peak coordinates. Results are corrected for multiple comparisons.

Region	MNI coordinates (X Y Z)			Cluster extent (voxels)	t-value
L inferior temporal sulcus / posterior fusiform gyrus	−44	−46	−16	125	3.95
L superior temporal gyrus	−64	−22	4	479	4.96
	−40	−14	−8		4.08
R amygdala	32	−6	−12	234	6.67
L amygdala	−20	−4	−6	67	4.43
R superior temporal gyrus	66	−26	6	502	4.96
	48	0	−8		3.83

semantic distance would appear. Hence, a whole-brain analysis was performed to identify areas where (after multiple-comparison correction) higher surprisal or semantic distance resulted in more activity compared to the reversed-speech baseline. Single-subject statistical maps were estimated using the General Linear Model as implemented in SPM8, and subsequent group analysis involved testing every voxel's significance over participants against zero ('Random-effects analysis'). Single-subject maps for the semantic distance and surprisal regressors tested for positive relationships with neural activity, and compared whether these positive relationships were stronger during listening to the stories than the reversed speech baseline. We used a significance threshold of $p < .005$ with 54-voxel extent to correct for multiple comparisons. The correction was based on a large number (10,000) of simulations estimating the critical number of voxels per region to arrive by chance (Slotnick, Moo, Segal, & Hart Jr., 2003).

For surprisal, the relevant areas corresponded largely to what was found before by Willems et al. (in press). In particular, the left inferior temporal sulcus / posterior fusiform gyrus ('visual word form area'), bilateral posterior superior temporal gyrus, and the amygdala bilaterally were found to be sensitive to word surprisal (see Figures 4 and 5 and Table 1). Willems et al. (in press) interpreted activation of the visual word form area as an indication that probability-based word prediction goes all the way down to word-form prediction.

Using the 'narrow' criterion for semantic distance, there are no areas where the measure significantly predicts brain activity (Figure 4). Under the 'wide' criterion, hower, areas that become more active on words with larger semantic distance are the left anterior temporal pole and anterior middle temporal sulcus, the left posterior middle temporal gyrus, the precuneus, and the angular gyrus bilaterally (see Figure 5 and Table 2).

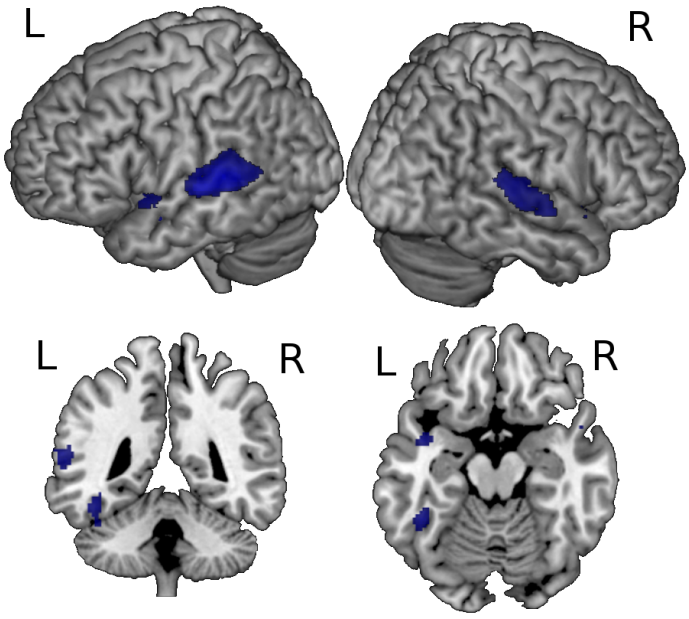

Figure 4: Brain areas that become significantly more active ($p < .05$ corrected for multiple comparisons) in response to larger surprisal. There is no significant effect of 'narrow criterion' semantic distance.

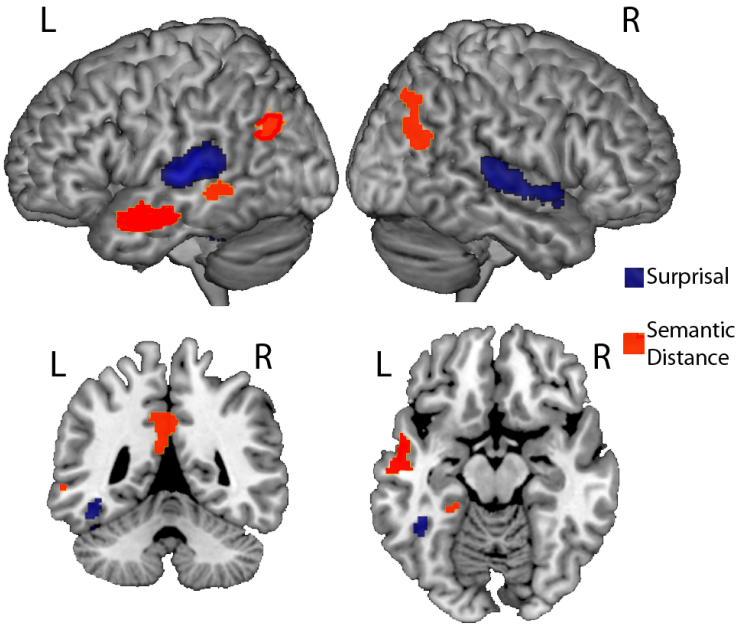

Figure 5: Brain areas that become significantly more active ($p < .05$ corrected for multiple comparisons) in response to larger surprisal (blue) or 'wide criterion' semantic distance (red).

Table 2: Brain areas that become significantly more active in response to larger 'wide criterion' semantic distance. The table displays a description of the region, coordinates in stereotaxic MNI space, the extent of the activation cluster, and the t-value of the reported voxels. Large clusters are represented by two peak coordinates. Results are corrected for multiple comparisons.

Region	MNI coordinates (X Y Z)			Cluster extent (voxels)	t-value
L anterior temporal pole / anterior	−58	−8	−18	363	4.65
middle temporal sulcus	−48	−10	−18		3.27
L posterior middle temporal gyrus	−62	−38	−4	89	3.68
Precuneus	0	−42	36	975	4.52
L angular gyrus	−40	−70	32	252	4.15
R angular gyrus	56	−68	24	180	4.33

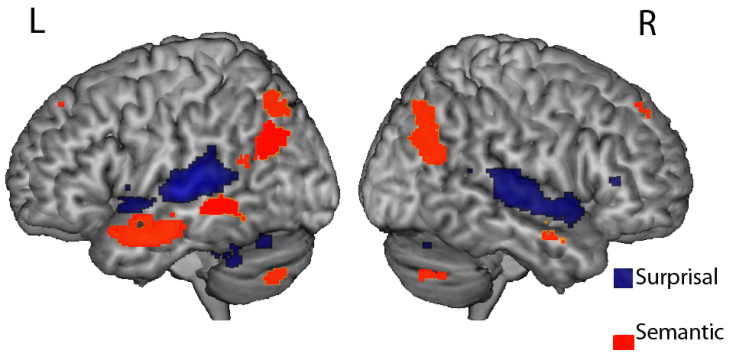

Figure 6: Brain areas that become significantly more active in response to larger surprisal (blue) or 'wide context' semantic distance (red), without multiple comparison correction ($p < .01$ voxel-wise significance threshold).

From the image in Figure 5, distinct areas appear to be activated by surprisal and semantic distance. However, the plotted activation map is thresholded: Activation that does not reach significance is not displayed. When an area is significantly activated by one measure but not by the other, this does not imply that the *difference* in activation due to the measures is significant because the two measures can be very close on either side of the threshold. However, when the threshold is decreased to $p < .01$ without multiple-comparision correction, the areas activated by the two measures remain distinct (Figure 6).

4 Discussion

Two processes can be identified by which words can become anticipated: top-down prediction using knowledge of syntagmatic relations in the language and bottom-up semantic priming due to paradigmatic relations

between words. The fact that we can conceive of these two separate pathways does not mean that the cognitive system uses them during language comprehension. Although effects of word predictability, as formalized by surprisal, are robust and well established, semantic priming effects on sentence comprehension appear to be more fickle (Camblin et al., 2007).

We took surprisal and semantic distance measures on all content words from natural sentences and texts, computed by models trained on a large text corpus. The n-gram model does not include any notion of lexical semantics so the surprisal values it estimates do not capture semantic relatedness. Conversely, the cosine distances between word vectors from the neural network model do not depend on word order or any other syntagmatic relation between words, so they do not incorporate word predictability (as we define it).

A comparison between the surprisal and distance measures to brain activation during language comprehension revealed first of all that each measure has significant effects over and above the other, at least when computing semantic distance between the current word and the previous context beyond the immediately preceding four words (i.e., using the 'wide' inclusion criterion). Under the common assumption that a weaker BOLD signal or N400 is indicative of reduced processing effort, we found that words that are more predictable or more strongly related to earlier words are easier to access or process. That is, they are anticipated to a greater extent. These surprisal and semantic distance effects generalize to different languages (English and Dutch), brain imaging methods (EEG and fMRI), stimuli types (individual sentences and narratives), and modalities (written and spoken stimuli). One may argue that the neural effects are actually caused by differences in integration difficulty rather than anticipatory processing, but this merely begs the question: Why would formal measures of word predictability and semantic relatedness be correlated to integration difficulty if not because they are measures of word anticipation?

Higher surprisal resulted in increased activity in bilateral posterior superior temporal areas, as well as the putative 'visual word form area' (VWFA), overlapping with what was found by Willems et al. (in press) on the same data set but using a different language model. The fact that surprisal, but not semantic distance, is related to VWFA activity suggests that our distance measure is not a word-unexpectedness measure in disguise: Prediction of upcoming words goes all the way down to pre-activating word form, but semantic priming apparently does not.

Words with larger semantic distance to previous words resulted in higher activation in, among others, the left temporal pole, angular gyrus and precuneus. In a meta-analysis by Binder, Desai, Graves, and Conant (2009) these areas were identified as parts of the lexical semantic system, which is consistent with our claim that the semantic distance measure quantifies semantic priming. Wehbe, Murphy, et al. (2014) recently found angular gyrus activity to correlate with semantic features from a distributional semantics model, again demonstrating the importance of this region to lexical semantics. Crucially, brain areas related to surprisal and to distance did not overlap, not even within the left temporal cortex, so even if we refrain from functional interpretation we can conclude that prediction and priming are neurally (and, therefore, most likely also cognitively) distinct processes. Both pathways to word anticipation are used during language

comprehension.

The 'narrow criterion' semantic distance measures did not show any significant effects in the fMRI study, although its EEG effect was almost as strong as that of surprisal (and of the 'wide criterion' alternative), indicating that the measure itself is viable. This discrepancy between fMRI and EEG results is likely to be due to the difference in stimuli modality, in particular the much faster presentation rate for the auditory fMRI study compared to the visual EEG study. In a lexical decision EEG experiment, Anderson and Holcomb (1995) found that in auditory presentation, no semantic priming effect (indicated by reduced N400 for related versus unrelated prime) occurred with a 200 ms stimulus-onset asynchrony (SOA) whereas an 800 ms SOA did suffice. Visually presented stimuli showed priming effects for both SOAs. Average word duration in the fMRI study was approximately 200 ms and even shorter for function words as these tend to be short. Consequently, if two content words are separated by one or two function words, there may not be enough time for a priming effect to appear. The fact that surprisal effects *are* visible at such short duration again shows that prediction and priming are indeed distinct neural processes.

Turning now to the EEG results, we see that surprisal and semantic distance have identically timed ERP effects, at least to the extent they relate to the N400. It is quite remarkable that the two measures have the same effect size, in the sense that one standard deviation increase in either measure resulted in the same amount of increase in N400 size, considering that predictability effects are much more robust than priming effects, at least in reading times (Camblin et al., 2007).

Camblin et al. (2007) found that N400 effects of semantic congruency (which, we have argued, is a form of predictability) precede those of priming. We did not replicate this but their study used short pieces of cohesive discourse which may have resulted in stronger predictions than our individual sentences. A time-separation between congruency and priming effects is of course fully consistent with our claim that surprisal and semantic distance are neurally distinguishable.

In addition to the N400 effects, Figure 3 provides some evidence for an early frontal positivity, possibly a P2 component, which is sensitive to surprisal first and to semantic distance later. Keeping in mind that we cannot make strong claims about the reliability of this effect as we did not plan to look at other components than N400, it does match Camblin et al.'s (2007) finding of priming effects arising later than congruency effects.

5 Conclusion

We showed that surprisal and semantic distance can have neurally distinguishable effects during language comprehension. We take surprisal as measure of *prediction* during language comprehension whereas semantic distance quantifies *priming*. In the N400 response the effects of surprisal and semantic distance was not distinguishable, whereas the fMRI results did show separate neural correlates for the two measures. While these measures are probably both important for language comprehension, we show that they have their

effect via separable neural correlates. This calls for current models of sentence comprehension to explicitly take priming and prediction into account as two separate mechanisms, both involved in anticipation during language comprehension.

In neuroimaging studies of language, there is a recent trend towards the use of naturalistic stimuli as opposed to hand-crafted experimental items (e.g., Wehbe, Murphy, et al., 2014; Willems, 2015). Making use of natural variation in language, rather than imposing extremes such as semantic anomalies or syntactic violations, increases generalizability of the results and reduces the risk of artifacts, for example caused by participants adjusting their processing strategies to the nature of the stimuli. In addition, it has the advantage that very rich data sets can be collected, which allow for many different analyses. As a case in point, we opted for re-analyzing published fMRI and EEG data as a first test of the relation between brain activity and computational measures of semantic relatedness. Although the individual data sets suited our needs and the converging results are evidence that the findings generalize well, it can be considered a drawback that different stimuli sets were used for the fMRI and EEG studies. For example, we cannot be sure that identical timing of surprisal and distance effects holds up in spoken narratives as opposed to written individual sentences. Future work using MEG, where both location and timing can be determined with high resolution, may provide the answer to this question.

The use of naturalistic stimuli combines well with parametric designs where computational characterization of materials is compared to brain activity, because the models can quantify every word of the stimuli. This method has previously relied on probabilistic language models (Frank et al., 2015; Hale et al., 2015; Wehbe, Vaswani, et al., 2014; Willems et al., in press) and the current work is the first that also applies it to study the effect of semantic relatedness. Hence, our results further demonstrate the value of using computational models for prediction brain activity during comprehension of naturalistic stimuli.

Acknowledgments

The work presented here was funded by the European Union Seventh Framework Programme under grant number 334028 awarded to SLF, a Vidi grant from the Netherlands Organisation for Scientific Research (NWO) to RW, and by NWO Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Anderson, J. E., & Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology*, *32*, 177–190.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796.

Brakel, P., & Frank, S. L. (2009). Strong systematicity om sentence processing by simple recurrent networks. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1599–1604). Austin, TX: Cognitive Science Society.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of predictions and contextual support on lexical processing: prediction takes precedence. *Cognition*, *136*, 135–149.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*, 103–128.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Frank, S. L., 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. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11.

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Günther, F., Dudschig, C., & Kaup, B. (in press). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*.

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

Hale, J. T., Lutz, D., Luh, W., & 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* (pp. 89–97). Denver, Colorado: Association for Computational Linguistics.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite

holographic lexicon. *Psychological Review*, *114*, 1–37.

Kuperberg, G. R., & Jaeger, T. F. (in press). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*, 484–502.

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

Mesnil, G., He, X., Denk, L., & Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Proceedings of Interspeech.*

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*, 545–567.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop.*

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of Interspeech.*

Mitchell, J., & Lapata, M. (2009). Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 430–439). Singapore: Association for Computational Linguistics.

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.

Mitchell, T., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–1195.

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

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.

Poser, B. A., Koopmans, P. J., Witzel, T., Wald, L. L., & Barth, M. (2010). Three dimensional echo-planar imaging at 7 Tesla. *NeuroImage*, *51*, 261–266.

Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, *48*, 2172-2183.

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

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart Jr., J. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research*, *17*, 75–82.

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*, 157–168.

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

Stafura, J. Z., & Perfetti, C. A. (2014). Word-to-text integration: message level and lexical level influences in ERPs. *Neuropsychologia*, *64*, 41–53.

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

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, *8*, 485–531.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, *9*, e112575.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (p. 233-243). Doha, Qatar: Association for Computational Linguistics.

Willems, R. M. (Ed.). (2015). *Cognitive neuroscience of natural language use*. Cambridge, UK: Cambridge University Press.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (in press). Prediction during natural language comprehension. *Cerebral Cortex*.