Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response

Mehdi Parviz, Mark Johnson
Department of Computing
Macquarie University
Sydney, Australia
{mehdi.parviz,mark.johnson}
@mq.edu.au

Blake Johnson, Jon Brock
Macquarie Centre for Cognitive Science
Macquarie University
Sydney, Australia
{blake.johnson,jon.brock}
@mq.edu.au

Abstract

The N400 is a human neuroelectric response to semantic incongruity in on-line sentence processing, and implausibility in context has been identified as one of the factors that influence the size of the N400. In this paper we investigate whether predictors derived from Latent Semantic Analysis, language models, and Roark's parser are significant in modeling of the N400m (the neuromagnetic version of the N400). We also investigate significance of a novel pairwise-priming language model based on the IBM Model 1 translation model. Our experiments show that all the predictors are significant. Moreover, we show that predictors based on the 4-gram language model and the pairwise-priming language model are highly correlated with the manual annotation of contextual plausibility, suggesting that these predictors are capable of playing the same role as the manual annotations in prediction of the N400m response. We also show that the proposed predictors can be grouped into two clusters of significant predictors, suggesting that each cluster is capturing a different characteristic of the N400m response.

1 Introduction

There is increasing interest in using computational models to help understand on-line sentence processing in humans. New experimental techniques in psycholinguistics and neurolinguistics are producing rich data sets that are difficult to interpret using standard techniques, and it is reasonable to ask if the statistical models developed in computational linguistics can be helpful here (Keller, 2010).

The N400 is a human brain response to semantic incongruity or implausibility that has been widely studied in psycholinguistics and neurolinguistics. A large set of factors has been shown to influence the strength of the N400, including intra- and extra-sentential context (Kutas and Federmeier, 2000; Van Petten and Kutas, 1990). Here we study the strength of the N400 as measured by magnetoencephalography (MEG) (so the signal we study is sometimes called the N400m) on sentence-final words in a variety of "constraining" and "non-constraining" sentential contexts (Kalikow et al., 1977). For example, *Her entry should win first prize* is an example of a constraining-context sentence, while *We are speaking about the prize* is a non-constraining context sentence (target words are underlined in this paper).

This paper shows that language models of the kind developed in computational linguistics can be used to help identify the factors that determine the strength of the N400. We investigate a number of different kinds of predictors constructed from a variety of language models and Latent Semantic Analysis (LSA) to determine how well they describe the N400. The first set of predictors is derived from LSA, which is a method for analysing relationships between a set of documents and the terms they contain (Mitchell et al., 2010). LSA has been successfully applied in similar research areas such as eye-movements and word-by-word reading times. Our experiments show that these predictors are significant in modeling the N400m response. The second set of predictors is that proposed by Roark et al. (2009), which is derived from the Roark (2001)

Mehdi Parviz, Mark Johnson, Blake Johnson and Jon Brock. 2011. Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of Australasian Language Technology Association Workshop*, pages 38−46

parser and designed to be useful in psycholinguistic modeling. While one of these predictors is statistically significant (lexical entropy), we observe that many of the prime-target word pairs appearing in our experimental sentences do not appear in the 1-million word Wall Street Journal - Penn Treebank (WSJ-PTB) corpus that this parser is trained on, so this model cannot capture the association between these words. This leads us to experiment with language models trained on larger corpora.

Using the SRI-LM toolkit (Stolcke, 2002) we construct a 4-gram language model based on the Gigaword corpus (Graff et al., 2005), and show that predictors based on it are also statistically significant predictors of the N400m. However, we go on to observe that many of the prime-target word pairs in our experimental sentences are separated by more than 3 words, so there is no way that a 4-gram language can capture the relationship between these words.

This leads us to develop a "pairwise-priming" language model that captures longer-range dependencies between pairs of words. This pairwise priming model is based on the IBM Model 1 machine translation model (Brown et al., 1993), and trained using a similar EM-procedure. We train this model on Gigaword, and show that predictors based on this model are also statistically significant.

Finally, we compare the predictors from the various language models with the original manual classification of the experimental sentences into "constraining" or "non-constraining" contexts given by Kalikow et al. (1977). We show that the predictor based on LSA is statistically significant even when the human "constraining" annotations are present as a factor. We also find out that the 4-gram model and the pairwise-priming model are highly correlated with this manually-annotated context predictor. These findings suggest that the predictors can be grouped into two clusters i.e., one that contains the LSA predictor, and another one that contains the manually-annotated context predictor, the pairwise-priming predictor, the 4-gram language model predictor, and the lexical entropy predictor.

2 Related work

One recent strand of work uses machine-learning to perform "mind reading", i.e., predicting what a subject is seeing or thinking based on information about their neural state. Mitchell et al. (2008) have trained a classifier that identifies the word a subject is thinking about from input derived from fMRI images of the subject's brain, and Murphy et al. (2009) have constructed a similar classifier that takes EEG signals as its input. Abstractly then, this work uses classifiers that take as input information about a subject's brain state to predict the (linguistic or visual) stimulus the subject is exposed to.

A more traditional line of research tries to identify factors that cause particular psycholinguistic or neurolinguistic responses. For example, Hale (2001), Bicknell and Levy (2009) and many others show that predictors derived from on-line parsing models can help explain eye-movements and word-by-word reading times. Abstractly, this work involves building statistical models which take as input properties of the stimuli presented to the subject (i.e., the sentence they are hearing or reading) to predict their psychological or neural responses. The goal of this line of research is to establish which properties of the input sentence or the parsing model's state determine the psychological or neural responses, rather than just predicting these responses as accurately as possible.

The work that is perhaps most closely related to this paper is by Bachrach (2008), who tries to identify which factors are responsible for specific activation patterns in fMRI brain images of subjects reading natural texts. He found that predictors derived from the Roark (2001) parser were most explanatory. Roark et al. (2009) have subsequently identified a number of such predictors; we investigate these in our analysis below.

3 Experimental data

The N400 is a component of time-locked EEG signals known as *event-related potentials* (ERP) that occurs in sentences containing semantically unexpected or anomalous words (Kutas and Hillyard, 1980). It is so-called because it is a negative-going deflection that peaks around 400 milliseconds post-stimulus onset. There has been considerable research into the factors that influence the strength of the N400. Inverse word frequency and contextual unpredictability (e.g., as quantified by Cloze prob-

ability) are both significant predictors of the N400 (Van Petten and Kutas, 1990). The strength of the N400 is sometimes taken to be a measure of the "effort" required for "semantic integration" in on-line sentence processing.

For example, there is a much stronger N400 at the target word *building* in the sentence *a sparrow is a kind of building* than there is at the word *bird* in *A sparrow is a kind of bird*. Interestingly, while the N400 is sensitive to the global context in which the target word is located, the N400 does not seem to be directly sensitive to the truth conditions of the sentence (Kutas and Federmeier, 2000). Thus sentential negation does not seem to directly affect the strength of the N400. For example, a strong N400 occurs in *A sparrow is not a kind of building*, as compared to *A sparrow is not a kind of bird*.[1] This observation inspired the pairwise-priming model discussed below.

As previously mentioned, N400s are usually studied using EEG. In this work we use magnetoencephalography (MEG) to study the N400; the signal we analyse here is sometimes called the N400m to indicate its provenance. We used MEG because this study is the first step in a project to use statistical models to study the neural mechanisms involved in language processing, and MEG seems ideally suited to this work.

MEG is a non-invasive technique for imaging electrical activity in the brain by measuring the magnetic fields it produces using arrays of SQUIDs (superconducting quantum interference devices). It has a number of potential advantages over competing technologies such as fMRI and EEG. For example, MEG has a much faster response latency than fMRI because MEG directly measures electrical activity while fMRI measures the hemodynamic response caused by that activity. Because magnetic fields are less distorted than electric fields by the scalp and the skull, MEG has a better spatial resolution than EEG, which should help us localise neural processes more accurately.

However in this first study we do not exploit these advantages of MEG, but just average the signals collected by 12 MEG sensors over a time window containing the target word. This produces a single numeric value for each trial which we call the N400m, which we model below.

Stimuli consisted of 180 sentences drawn from the list published by Kalikow et al. (1977) and synthesized using TextAlound (NextUp, Clemmons, NC). They were presented to 22 listeners via insert earphones (Etymotic Research Inc. Model ER-30, Elk Grove Village, IL). There were 90 examples of "constraining context" sentences, i.e., with predictable endings (e.g. *He got drunk in the local bar*) and 90 examples of "non-constraining context" sentences, i.e., with unpredictable endings (e.g. *He hopes Tom asked about the bar*). Each target word appears both in a constraining context sentence and in a non-constraining context sentence. To maintain vigilance during the experiment, there were 10 catch trials consisting of sentences containing the word *mouse*, where subjects were required to press a button. The three types of sentences were presented in randomized order.

MEG amplitudes were extracted from a cluster of 12 sensors over the left hemisphere where the largest N400m responses were obtained over subjects. Amplitudes in femto-Tesla were averaged over these sensors and over a time window of 400-600 ms. MEG data was digitized with a sample rate of 1000 Hz and were filtered offline with a bandpass of 0.1 to 40 Hz. Data was epoched relative to the onset of the terminal word of each sentence using a 1200 ms window (-200 to 1000 ms).

4 Hypothesis-testing

Our goal in this paper is to identify the factors that significantly influence the N400m, rather than predicting the N400m responses as accurately as possible. We use statistical methods for hypothesis testing (e.g., likelihood ratio tests) to do this. The next two paragraphs explain why we use these methods rather than the held-out test set methodology usually used in computational linguistics.

The goal of most statistical modeling in computational linguistics is prediction, which in turn involves generalisation to previously-unseen contexts, and the held-out test set methodology measures the ability of a model to generalise correctly. One might attempt to identify significant predictors by build-

[1] The fact that the conditional probability of a word in a sentence does not depend on that sentence's veracity may be relevant here.

ing the best machine learning model of the N400m one can, and see which features that model incorporates. However, many state-of-the-art machine learning methods are capable of exploiting very large sets of possibly redundant features and control over-learning via regularisation. The fact that such a method includes a particular predictor as a feature does not mean that this predictor is significant; e.g., the method may assign the feature a very small (but non-zero) weight. Intuitively, the goal of a machine-learning method is to make the most accurate prediction possible, not to identify the significant predictors.

Instead, we formulate the problem as one of *hypothesis testing*. The statistical techniques used to do this involve the construction of linear models similar to those used in some machine-learning methods, but they also permit us to perform hypothesis testing and posterior inference. For example, by computing confidence intervals on a predictor's weight in such a model we can see whether that confidence interval contains zero, and hence whether the predictor is significant. We also use likelihood-ratio tests below to assess the significance of predictors.

We used a quantile plot to identify outliers in the N400m data; four responses were removed, and one response value was unavailable, producing five missing values for the N400m in total. The N400m data range from -1,054 to 1,362 with a mean of 14, a variance of 172 and an interquartile range of (-68,100). We normalised the N400m responses by subtracting the per-subject mean and then dividing by the per-subject standard deviation. The N400m responses are the values of the `Response` variable in the models below.

4.1 Parser-based predictors

The Roark (2001) parser is an incremental syntactic parser based language model that uses rich lexical and syntactic contexts as features to predict its next moves. It uses a beam search to explore the space of partial parse trees. Bachrach (2008) found that predictors derived from the incremental state of the Roark parser were highly significant in models of their fMRI data; this work motivated us to explore predictors like lexical entropy and lexical surprisal based on the Roark parser here. Roark et al. (2009) describes in detail how a variety of predictors can be extracted from the Roark parser. We used Roark's parser to compute these predictors for the target words in all 180 of the experimental sentences used here.

4.2 4-gram language model predictors

We used the Gigaword corpus which contains 1.5 billion words in 82 million sentences (Graff et al., 2005). We trained a 4-gram language model with Kneser-Ney smoothing and unigram caching using the SRI-LM toolkit (Stolcke, 2002). We used this language model to estimate the conditional probabilities of the target words given the words in their preceding context in all of the experimental sentences. These probabilities are often very close to zero, can vary by many orders of magnitude, and may be highly skewed. In order to mitigate the effect of these properties we used log ratio of these probabilities to the unigram probabilities of the target words as predictors. This is called the `P4` predictor below.

4.3 Pairwise-priming predictors

By definition, a 4-gram language model only captures dependencies between words within a 4 word window. However, many of the experimental sentences contain dependencies between words that are more than 3 words apart. For example, in "constraining context" sentences such as *The steamship left on a cruise* or *We camped out in our tent*, the priming words *steamship* and *camped* do not appear in the 4 word window containing the target words *cruise* and *tent*, but these priming words are intuitively responsible for making the corresponding target words more likely.

It is plausible that "trigger" language models can capture these kinds of longer-range dependencies (Goodman, 2001). There are a wide variety of such models, and it would be interesting to see which of them are most useful for constructing N400 predictors. Rather than using an existing trigger language model, we develop our own "pairwise-priming" language model here. This model is especially designed to identify longer-range interactions between pairs of words, which we believe is consistent with the description given by Kutas and Federmeier (2000) of the factors influencing the strength of the N400. This model is also especially simple to estimate us-

ing a variant of the EM training procedure for IBM Model 1.

The model is a simple additive mixture model. Each word w_i in a sentence is associated with a *context* C_i which is used to predict w_i. The context C_i is a bag containing the words that precede w_i in the sentence and that also belong to a 60,000 word vocabulary W, plus 5 instances of a special *null word* token.[2] The vocabulary consists of the most frequent words in the Gigaword corpus, from which 60 open-class stop words have been removed. Our model is parameterised by a matrix $\boldsymbol{\theta}$, where $\theta_{w_i|w_j}$ is the probability of generating w_i given that w_j is in the context C_i. The probability $\mathrm{P}(w_i \mid C_i)$ of generating w_i in the context C_i is approximated by an additive mixture:

$$\mathrm{P}(w_i \mid C_i) \quad = \quad \frac{1}{|C_i|} \sum_{w_j \in C_i} \theta_{w_i|w_j}.$$

This is a conventional generative model in which each word w_i is generated from the words in its context C_i, and it is straightforward to estimate the pairwise-priming parameters $\boldsymbol{\theta}$ using a variant of the IBM Model 1 EM training procedure. This EM procedure computes a sequence of estimates $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ that approximate the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. The M-step computes $\boldsymbol{\theta}^{(t+1)}$ from the expected pairwise counts obtained using $\boldsymbol{\theta}^{(t)}$:

$$\theta^{(t+1)}_{w'|w} \quad = \quad \frac{\mathrm{E}_{\boldsymbol{\theta}^{(t)}}[n_{w',w}]}{\sum_{w'' \in \mathcal{W}} \mathrm{E}_{\boldsymbol{\theta}^{(t)}}[n_{w'',w}]}.$$

The E-step calculates the expected counts $\mathrm{E}_{\boldsymbol{\theta}^{(t)}}[n_{w',w}]$ given the current parameters $\boldsymbol{\theta}^{(t)}$:

$$\mathrm{E}_{\boldsymbol{\theta}^{(t)}}[n_{w',w}] \quad = \quad \sum_{\substack{i\,:\,w_i = w' \\ j\,:\,w_j = w, w_j \in C_i}} \frac{\theta^{(t)}_{w'|w}}{\sum_{w'' \in C_i} \theta^{(t)}_{w'|w''}}$$

In the E-step we skip the first four w_i words of every sentence because we think their contexts C_i are likely to be too small to be useful, but we did no experiments to test this. We initialised with the uniform distribution (by using an argument analogous to the one for IBM model 1 it is easy to show the log-likelihood surface is convex), and ran 10 EM iterations on the Gigaword corpus to estimate $\widehat{\boldsymbol{\theta}}$.

Just as for the 4-gram models, we used the pairwise priming model to compute the conditional probability of the target words in the experimental sentences. Like the 4-gram models, we used log ratio of these probabilities to the probabilities of the target words as predictors. This is called the `Pq` predictor below.

4.4 Latent semantic analysis predictors

Another predictor used in applications such as modeling eye-movements and word-by-word reading times, is Latent Semantic Analysis (LSA). The basic idea of the LSA model is to create a "meaning representation" for words from a term-document co-occurrence matrix. Here we construct the model based on the co-occurrence of vocabulary and content-bearing words in a fixed-sized window of the Gigaword corpus (Graff et al., 2005). We used the 2,000 most-frequent words in the corpus as the content words and the 50,000 most-frequent words as the vocabulary. Each row in the matrix represents a vocabulary word, each column represents a content word, and each entry is the co-occurrence count $n_{i,j}$ of the ith vocabulary word and the jth content word within a window with 15 words length. The co-occurrence counts are normalised by dividing each $n_{i,j}$ by the sum of all the counts in the corresponding column:

$$w_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

LSA performs dimensionality reduction using Singular Value Decomposition (SVD). In order to retain 99% of the total variance, we used 795 right eigenvectors of the normalised co-occurrence matrix. Following Mitchell et al. (2010), we used the LSA model to generate a numerical value indicating the "similarity" of the target word to the words in its preceding context as follows: Let W_1, W_2, $\cdots$, W_n denote vectors representing the context words and let W_t denote a vector representing the target in

[2] The null word token plays the same role here as it does in the IBM Model 1 machine translation model (Brown et al., 1993). Moore (2004) points out that including multiple null word tokens reduces the tendency of the IBM Model 1 to find spurious low-frequency associations; we found here that while including multiple null word tokens in the C_i is important, the results do not depend strongly on the number of null word tokens used.

a given sentence.

$$\begin{matrix} W_1 & W_2 & \cdots & W_n & W_t \end{matrix}$$
$$\begin{bmatrix} w_{1,1} \\ w_{1,2} \\ \vdots \\ w_{1,m} \end{bmatrix} \begin{bmatrix} w_{2,1} \\ w_{2,2} \\ \vdots \\ w_{2,m} \end{bmatrix} \cdots \begin{bmatrix} w_{n,1} \\ w_{n,2} \\ \vdots \\ w_{n,m} \end{bmatrix} \begin{bmatrix} w_{t,1} \\ w_{t,2} \\ \vdots \\ w_{t,m} \end{bmatrix}$$

We multiply the context-word vectors element-wise to produce a single vector H representing the context as follows:

$$h_i = \prod_{j=1}^{n} w_{j,i}$$

Then the similarity of a target word to the context words is given by the cosine of the angle between H and W_t, i.e.:

$$\text{sim}(H, W_t) = \frac{H^T W_t}{\|H\| \|W_t\|}$$

We call $\text{sim}(H, W_t)$ the `LSA` predictor below.

5 Experimental Results

We normalised the N400m responses by subtracting the per-subject mean and then dividing by the per-subject standard deviation. Similarly, we normalised the values of predictors. We used the non-linear regression package mgcv v1.7-6 (Wood, 2006; Wood, 2011) distributed with the *R* statistical environment to predict the N400m response. We used the manually-annotated context predictor (`Context`) as a linear parametric predictor, and all the other types of predictors i.e., the 4-gram language model predictor (`P4`), the pairwise priming predictor (`Pq`), the LSA predictor, and the predictors based on Roark's parser, as penalized cubic spline functions (up to 20 degrees of freedom).

5.1 Models with one predictor

We first start with models with one predictor to find out which predictors are significant. Table 1 lists the significant predictors, where significance is determined by a likelihood ratio test. Of all the predictors described by Roark et al. (2009) only the `LexH` predictor (lexical entropy) is a significant predictor according to a likelihood-ratio test. Perhaps it should not be surprising that lexical entropy strongly predicts the N400m response; the lexical entropy is a measure of the predictive uncertainty of the target word, and the N400 is strongest in less predictive contexts.

Predictor	Df	p-value
Context	1	1.53e-11 ***
Pq	2.3479	4.84e-10 ***
P4	2.067	5.30e-10 ***
LexH	3.2197	1.75e-04 ***
LSA	1.6707	5.28e-04 ***

Table 1: P-values and degrees of freedom as determined by likelihood ratio test for non-linear regression models with only one predictor

	Context	Pq	P4	LexH
Pq	-0.76***			
P4	-0.76***	0.96***		
LexH	0.41***	-0.38***	-0.38***	
LSA	-0.15*	0.09	0.10	-0.06

Table 2: Correlation matrix of different types of predictor

5.2 Combining predictors

In this section, we combine all the predictors to create a single model. From the correlation matrix of the predictors (Table 2), we can see that some of these predictors are highly correlated. Not surprisingly, when we combined all the predictors we discovered that some of predictors are redundant. We performed backwards selection using p-values to drop insignificant predictors (Wood, 2011). In backwards selection, first we construct a model with all the predictors, then we drop the single predictor with the highest non-significant p-value from the model. We repeat re-fitting, dropping insignificant predictors until all remaining predictors are significant. The results of performing backwards selection show that only the manually-annotated context predictor and the LSA predictor are significant (Table 3):

`Response` $\sim$ `Context + LSA`

In order to construct a model without the manually-annotated context predictor, we removed the manually-annotated context predictor from the model and re-performed backwards selection. The

Predictor	Df	p-value
Context	1	2.34e-10 ***
LSA	2.779	0.0186 *

Table 3: P-values and degrees of freedom of the predictors in the combined model after performing backwards selection

Predictor	Df	p-value
LSA	3.165	0.00405 **
Pq	1.987	0.01817 *
P4	2.158	0.04340 *

Table 4: P-values and degrees of freedom of the predictors in the combined model without the manually-annotated context predictor after performing backwards selection

results show that the combination of the pairwise priming predictor, the 4-gram language model predictor, and the LSA predictor are significant (Table 4):

`Response ~ LSA + Pq + P4`

In order to minimise the effect of collinearity of predictors, we applied PCA to find principal components of the predictors' space. In Table 5, the matrix of eigenvectors is shown. Treating the principal components as predictors, we performed backwards selection to find a set of significant principal components. In Table 6 the p-values of all the principal components are presented. After performing backwards selection, only the first two principal components are significant (Table 7):

`Response ~ PC1 + PC2`

As can be seen, in the first principal component (PC1) `Context`, `Pq` and `P4` are dominant, while in the second principal component `LSA` is dominant. We can conclude that proposed predictors can be grouped into two clusters; one that contains the LSA predictor, and another that contains the manually-annotated context predictor, the pairwise-priming predictor, the 4-gram language model predictor, and the lexical entropy predictor.

Hierarchical clustering also suggests that the set of predictors cluster into two groups. Figure 1 depicts a hierarchical clustering of the predictors based on Spearman's rank correlation (Myers and Well, 2003). As this figure shows, the similarity between the LSA predictor and other predictors is close to zero.

	PC1	PC2	PC3	PC4	PC5
Context	0.52	-0.01	0.10	0.85	0.01
Pq	-0.55	-0.09	-0.24	0.36	0.71
P4	-0.55	-0.07	-0.24	0.37	-0.71
LexH	0.33	0.05	-0.94	-0.10	-0.00
LSA	-0.10	0.99	0.01	0.07	0.01
Eigenvalue	2.92	0.98	0.76	0.29	0.04

Table 5: The principal components of the predictors' correlation matrix

	Df	p-value
PC1	1.000	2.23e-10 ***
PC2	5.230	0.00627 **
PC3	1.445	0.85781
PC4	1.000	0.59922
PC5	2.412	0.21918

Table 6: P-values and degrees of freedom for the principal components in the combined model before performing backwards selection

6 Conclusions and future work

This paper has studied a variety of predictors of the N400m response derived from an incremental parsing model (Roark et al., 2009), from Latent Semantic Analysis, and from two language models trained on the Gigaword corpus (Graff et al., 2005). We found that many of the predictors derived from these models were significant, suggesting that these kinds of models may be useful for understanding the N400m response. We also examined combining predictors to build a single model.

We can summarize our results as follows:

- A wide range of predictors are significant predictors of the N400m response on their own:
  -the manually-annotated context predictor, `Context`
  -the LSA predictor, `LSA`
  -the lexical entropy predictor, `LexH`, based on Roark's parsing model
  -the 4-gram language model predictor, `P4`, and
  -the pairwise-priming predictor, `Pq`
- These predictors can be grouped into two clusters:

	Df	p-value
PC1	1.188	5.78e-14 ***
PC2	4.848	0.0052 **

Table 7: P-values and degrees of freedom for the principal components in the combined model after performing backwards selection

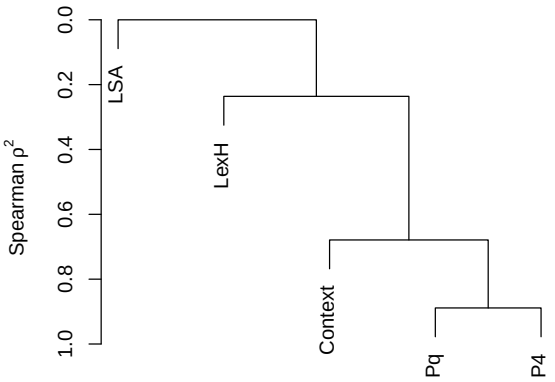

Figure 1: Hierarchical clustering of predictors, using square of Spearman's rank correlation as similarity measure

-1: The manually-annotated context predictor (`Context`), the 4-gram language model predictor (`P4`), the pairwise-priming predictor (`Pq`), and the lexical entropy (`LexH`), and
-2: The Latent Semantic Analysis predictor (`LSA`)

This latter result suggests that these two groups of predictors are capturing separate factors of the N400m response. Of course this work just scratches the surface in terms of possible applications of statistical language models to neurolinguistics. Clearly it would be interesting to apply a much wider variety of statistical models to the N400 data. Perhaps parsing models would do better if they could be trained on Gigaword-sized corpora. As we noted above, MEG is capable of producing rich temporal and spatial information about neural processes, presenting new opportunities for using statistical language models to help understand how language is instantiated in the human brain.

References

Asaf Bachrach. 2008. *Imaging Neural Correlates of Syntactic Complexity in a Naturalistic Context*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Klinton Bicknell and Roger Levy. 2009. A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 665–673, Boulder, Colorado, June. Association for Computational Linguistics.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Joshua Goodman. 2001. A bit of progress in language modeling. *Computer Speech and Language*, 14:403–434.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2005. *English Gigaword Second Edition*. Linguistic Data Consortium, Philadelphia.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166. Association for Computational Linguistics.

D. N. Kalikow, K. N. Stevens, and L. L. Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal ofthe Acoustical Society of America*, 61(5):1337–1351.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden, July. Association for Computational Linguistics.

Marta Kutas and Kara D. Federmeier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science*, 4(12):463–470.

Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207:203–208.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A.

Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *ACL*, pages 196–206.

Robert C. Moore. 2004. Improving IBM word-alignment Model 1. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brian Murphy, Marco Baroni, and Massimo Poesio. 2009. EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 619–627, Singapore, August. Association for Computational Linguistics.

Jerome L. Myers and Arnold D. Well. 2003. *Research Design and Statistical Analysis (second edition ed.)*. Lawrence Erlbaum.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 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*, pages 324–333, Singapore, August. Association for Computational Linguistics.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Cyma Van Petten and Marta Kutas. 1990. Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, 18(4):380–393.

S.N Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

S. N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.