Prediction of Word Forms During Language Comprehension

Leif Oines

University of Colorado at Boulder, leifdoines@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/psyc_gradetds

Part of the Cognitive Psychology Commons, Linguistics Commons, and the Neurosciences Commons

Recommended Citation

https://scholar.colorado.edu/psyc_gradetds/149

This Dissertation is brought to you for free and open access by Psychology and Neuroscience at CU Scholar. It has been accepted for inclusion in Psychology and Neuroscience Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
PREDICTION OF WORD FORMS DURING LANGUAGE COMPREHENSION

by

LEIF OINES

B.A., University of California Santa Cruz, 2009

M.A., University of Colorado Boulder, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Psychology and Neuroscience

2017
This thesis entitled:
Prediction of Word Forms During Language Comprehension
written by Leif Oines
has been approved for the Department of Psychology and Neuroscience

______________________________
Albert Kim (Chair)

______________________________
Matthew Jones

Date________________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol # 15-0802
In conversation or during reading, we sometimes find ourselves making predictions about the identity of an upcoming word or phrase. This phenomenon has been reflected in the results of laboratory experiments that show changes in eye gaze patterns or the Electroencephalogram (EEG) prior to encountering a predicted word. However, questions remain about exactly what is predicted during language comprehension, as well as how often predictions are computed. Existing results most clearly support the prediction of upcoming words’ meanings, while lower-level predictions of a word’s perceptual features are less well supported. Furthermore, most evidence for prediction has come from tasks where participants read language that is designed to be predictable, so the importance of predictive processes for typical language use also remains unclear. Meanwhile, increasing attention has been paid to more general models of brain function that posit prediction and prediction error as representing the two fundamental “units of account” that are used in the passing of information between levels of the neural hierarchy. These predictive coding models imply that predictions during language comprehension must be generated constantly, and at all levels of representation. In this dissertation I describe three experiments that are designed to address the empirical matter of whether low-level word-forms are in fact predicted during comprehension, as predictive coding models demand. In study 1, I show that the unexpected omission of a highly predictable, sentence-embedded visual word still leads to a brain response that reflects the omitted word’s visual length. In study 2 I describe an experiment that suggests study 1’s findings reflect prediction error, rather than word-form prediction itself. Finally, study 3 provides evidence that word-form predictions are not limited to highly predictable
contexts. By providing consistent evidence for the predictions of word forms during comprehension, the results suggest prediction during language is not limited to only the most abstract levels of representation, or only the most constraining linguistic contexts. Instead, they suggest the neural machinery underlying word recognition is fundamentally predictive, confirming the predictions of predictive coding theories.
CONTENTS

1. Introduction..................................................................................................................................................1
   1.1. The evidence that comprehension is predictive
      1.1.1. Prediction of lexical identity............................................................................................................7
      1.1.2. Prediction of word forms....................................................................................................................10
   1.2. General models of prediction in perception and cognition.................................................................13
   1.3. Early Brain responses to visual words and their potential link to prediction.................................19
   1.4. The current studies................................................................................................................................22

2. Study 1: Early EEG Responses Betray the Forms of Words that Were Predicted but Never Seen
   2.1. Methods
      2.1.1. Participants.......................................................................................................................................24
      2.1.2. Stimuli...............................................................................................................................................24
      2.1.3. Procedure.........................................................................................................................................26
      2.1.4. Analysis of behavioral data.............................................................................................................27
      2.1.5. EEG data pre-processing..................................................................................................................28
      2.1.6. Analysis of P1 Event Related Potential...........................................................................................28
      2.1.7. Independent Components Analysis.................................................................................................30
   2.2. Results
      2.2.1. P1 ERP Results.................................................................................................................................31
      2.2.2. ICA Results.....................................................................................................................................35
   2.3. Study 1 Discussion
      2.3.1. The time-course of predictions........................................................................................................41
      2.3.2. What would a rational predictor predict? ..........................................................................................42
      2.3.3. Prediction or prediction error? ..........................................................................................................44

3. Study 2: Dissociating Prediction from Prediction Error
   3.1. Methods
      3.1.1. Participants and Stimuli....................................................................................................................46
      3.1.2. Procedure.........................................................................................................................................47
      3.1.3. Data analysis.....................................................................................................................................49
   3.2. Results.....................................................................................................................................................52
   3.3. Discussion and summary
      3.3.1. Effect of prediction error on the visual EEG response.................................................................57
      3.3.2. Timing of prediction, bottom-up processing and prediction error..............................................57

4. Study 3: Testing for consistent form-level prediction
   4.1. Methods
      4.1.1. LSTM model for word prediction........................................................................................................61
      4.1.2. GloVE Vectors for word representation.............................................................................................62
      4.1.3. Model architecture.............................................................................................................................63
      4.1.4. Model Training and Evaluation.........................................................................................................66
   4.2. Using the language model to predict visual EEG responses.............................................................67
   4.3. EEG data analysis
4.3.1. Replication on held-out dataset.................................................................71

4.4. Results
4.4.1. Analysis of EEG data related to Study 1 and Study 2 Non-target words............71
4.4.2. Replication of findings in held-out data.....................................................78

4.5. Summary & Discussion.......................................................................81

5. General Discussion
5.1. Prediction vs. Pre-activation.................................................................84
5.2. The utility of form prediction...............................................................86
5.3. The representational nature of form predictions.................................87
5.4. The relationship of early visual ERP components to word recognition........91
5.5. A predictive coding account of word recognition?.................................93

6. Conclusions and future directions.........................................................98

References...............................................................................................101
LIST OF TABLES

1. Table 1: Example stimulus sentences for study 1.................................................................25
2. Table 3.1: Results of regression analyses for study 2..........................................................54
LIST OF FIGURES

Note: Figures are numbered to reflect the study they are associated with.

1. Figure 1.1 ................................................................................................................................. 27
2. Figure 1.2 ................................................................................................................................. 29
3. Figure 1.3 ................................................................................................................................. 35
4. Figure 1.4 ................................................................................................................................. 36
5. Figure 1.5 .................................................................................................................................. 39
6. Figure 1.6 .................................................................................................................................. 40
7. Figure 2.1 .................................................................................................................................. 48
8. Figure 2.2 .................................................................................................................................. 54
9. Figure 2.3 .................................................................................................................................. 56
10. Figure 3.1 ................................................................................................................................ 64
11. Figure 3.2 ................................................................................................................................ 67
12. Figure 3.3 ................................................................................................................................ 73
13. Figure 3.4 ................................................................................................................................ 74
14. Figure 3.5a ................................................................................................................................ 75
15. Figure 3.5b ................................................................................................................................ 76
16. Figure 3.6 ................................................................................................................................ 79
17. Figure 3.7 ................................................................................................................................ 80
18. Figure D1 .................................................................................................................................. 89
19. Figure D2 .................................................................................................................................. 97
1. Introduction

Language comprehension is fundamentally sequential. As the meaning of an utterance or written passage unfolds over time, the neural apparatus of the person receiving the message must incrementally process each additional bit of information and update its representation of the discourse. This act of sequence modelling, done expertly by all fluent language users, relies on processes that decompose the incoming language sequence into a nested set of representations (e.g. phonemes, syllables and words) which simultaneously represent the input at a number of time-scales and semantic abstractions.

The serial nature of language stimuli, combined with the hierarchical representations that underpin their comprehension, naturally afford for a predictive system. Overlearned dependencies between elements of a linguistic sequence (be they phonemes, words, etc.) often provide enough information to allow confident and accurate prediction about properties of upcoming words. Furthermore, the hierarchical relationships between aspects of the linguistic representation (phrases depend on words, words depend on phonemes, etc.) provide a simple conceptual framework for the propagation of predictions through all levels of representation. For example, lexical-specific predictions can in turn prepare associated sensory (e.g. orthographic) representations in a top-down fashion.

Recently, a number of more general “predictive coding” models of perception and cognition (Rao & Ballard, 1999; Rao, 1999; Lee & Mumford, 2003; Friston & Seibel, 2009; Huang & Rao, 2011, Clark, 2013) have posited that prediction may be the exclusive purpose of the brain’s top-down connections. These frameworks are usually implemented as hierarchically generative bayesian models, in the sense that higher levels of representation actively generate hypotheses about what the representation should be at lower levels, and communicate these hypotheses via top-down connections. In turn, bottom-up connections convey the prediction error between the top-down predictions and each
layer’s current representation of what the input actually is. These models fundamentally differ from connectionist models of hierarchical feature detectors, in which stimuli are decoded in a more passive way from the bottom-up.

Thus, according to predictive coding models, predictions are passed through the entire extent of the representational hierarchy, including into the lowest sensory levels. In this dissertation I will argue that language comprehension represents the strongest test of this idea, as its underlying representations are deep in the sense that they stretch from grammatical and semantic categories at the most abstract, to the sub-phonemic and sub-orthographic features of perceptual form. A truly predictive coding architecture suggests that expectations generated at the highest levels will – as a basic feature of the system – be conveyed through all lower levels of the hierarchy, including to the most basic levels of form representation. The three studies I describe in this document represent rigorous tests of this idea. Providing evidence of prediction at the lowest perceptual levels of word recognition would serve to lend empirical teeth to what have so far been general and abstract models of perception and cognition. From another standpoint, the results also serve to inform the ongoing debate about prediction’s role in language specifically, which is very much an active concern for much of psycholinguistics.

Though it is still a matter of some debate, evidence has continued to accumulate about the predictive nature of language comprehension (see below for a brief overview). Much of this has confirmed what many already suspect from introspection – that we seem to have something of a mental “autocomplete” that sometimes anticipates the words of our fellow speakers, which we can in turn use when deciding when to jump into the conversation (Magyari & de Ruiter, 2012; Stivers et al., 2009). The extent to which linguistic prediction goes beyond these more transparent cases is still an open issue, however.
For one – although predictive coding models appear to require that they occur – it is far from clear that very low-level predictions would be computationally and/or metabolically beneficial in the long run. Indeed, most of the evidence for prediction during comprehension is at the lexical level, which does not require one to posit that the visual or auditory form of the word in-context be anticipated. Many factors would seem to stymie such a fully-formed prediction. For instance, a word can vary substantially in its perceptual features from speaker to speaker, or with changes in environmental conditions, such as in the lighting of the room in which you are reading, or in the level of ambient noise during your conversation. In general, there are reasons to believe that such low level predictions about an upcoming word’s perceptual features may on average be more prone to error, even when the prediction of the word’s actual identity turns out to be accurate. For this reason, it may be the case that predictions at lower levels may in fact be more conservative, occurring only when confidence in the perceptual input is high. Perhaps predictions at the absolute lowest, noisiest levels of representation are avoided entirely.

Furthermore, it is usually assumed that lexical-level predictions are made about particular words, rather than of multiple possible upcoming words at once. This conforms to the introspective autocomplete phenomenon described above – when we complete someone’s sentence in our head, we do not get the sense of simultaneously predicting multiple possible words. However, parallel, simultaneous prediction of many possible completions - perhaps in proportion to each completion’s likelihood – would make prediction more cost-effective on average, and may even make low-level prediction a viable strategy. This is easy to see – if we are predicting one particular word, but turn out to be wrong, then the cost will be proportional to the difference between our predicted word and the actual stimulus. However, if we hedge our bets by predicting a weighted combination of multiple possible completions, then the cost due to prediction error would be at least moderately reduced - even if only 1% of the eventual word’s features had figured in our prediction. Over the long run, such a
strategy could add up to a substantial metabolic savings, especially in contexts where a small handful of words are close to equiprobable. So far there has been scant evidence for such “weighted” predictions, a shortcoming that Study 3, reported below, was in part designed to address.

**Prediction from a Rational Perspective**

Prediction can be viewed from what is known as rational perspective (i.e., in the sense of Anderson, 1990), under the assumption that language comprehenders use all the information available to them when determining whether to compute a prediction (Levy, 2008; Kuperberg & Jaeger 2016). In these frameworks, language comprehenders maintain probability distributions over the potential meaning and structure of an utterance, and update these distributions as new information (e.g., a word) is added to the discourse. In particular, comprehension has been modeled as the act of shifting between a *prior* distribution, that represents the message’s meaning prior to encountering new words, to a *posterior* distribution, which represents the message’s meaning after each new word has been added to the discourse. The amount of information – or surprise – contributed by each new word can then be quantified as the Kullback-Leibler divergence (KLD) between the prior and posterior distribution.

Critically, prediction within these frameworks can be conceptualized as an anticipatory change to the prior distribution *before* the arrival of new words. Prediction may then be a useful strategy in proportion to the extent that such a change reduces (on average) the subsequent shift in beliefs once new information is actually encountered.

From the rational perspective, prediction will *always* makes sense, so long as all available information is brought to bear, and so long as (on average) it reduces the amount of work one must do while shifting from prior to posterior distributions. However, we might question the extent to which this latter assumption holds, at least with respect to low-level form prediction. In the rational framework, making no prediction at all is equivalent to maintaining a uniform prior distribution (equal probability
given to all possible outcomes). “Useful” prediction therefore entails anticipatory allocation of more probability mass to those possibilities that are closer to the actual outcome. Critically, however, if predictions are very far off, they can result in a larger surprise (e.g., larger KLD) than had no prediction been made at all. This may be a real hazard in some highly unpredictable contexts. From the most general perspective, then, we might outline three possible states of affairs with regards to form-level predictions. The first is the strong rational scenario, in which form predictions are always made, because on average there is enough information to successfully predict at least part of the stimulus. The second possibility is that low-level forms are never predicted, either because of what would be low rates of accuracy, or perhaps because of some other architectural constraint. And then there is a third possibility – that low-level predictions are opportunistic, or made only when the affordances for them are strong. Perhaps, for example, word forms are predicted only when there is enough evidence to commit to one particular outcome, or possible upcoming word.

Without some kind of additional constraint, I do not think that it is possible to determine which of these three outcomes is closest to the true state of affairs. In particular, the rational point of view cannot be comprehensively defended without showing that predictive changes to the prior distribution actually reduces its distance to the posterior on average. To put it in plainer terms, it is difficult to quantify whether the average context actually contains enough constraint to allow for better-than-chance prediction of upcoming word forms. Thus, rather than trying to resolve this question in theoretical terms, we designed three experimental studies that would provide empirical constraint on the issue of whether the language system is as predictive as some rational models suggest it should be.

In Study 1 we devise an experiment which overcomes the drawbacks of previous attempts to show top-down predictions of low-level word form features. This is done by carefully omitting some critical words during a reading task, and examining whether brain responses in these moments reflect
the form properties of the missing word. In Study 2 we address the interaction between prediction
generation and prediction error, asking whether the latter precedes the former in time, and whether
one or the other constitute the best explanation for the effects observed in Study 1. Finally, Study 3
examines whether evidence for form-level prediction can be found outside of highly predictable
linguistic contexts, and whether such predictions reflect probability-weighted combinations of many
outcomes, rather than simple commitments to one word.

The rest of the introduction is organized as follows. In 1.1 I provide some background on
previous experimental work concerning prediction during comprehension, including results that have
been explicitly interpreted as evidence for form-level prediction. In section 1.2 I provide some more
detail on general predictive coding models, paying particular attention to how they might be used to
account for the processes of word recognition. In 1.3 I provide some background on early, visually-
evoked Event Related Potentials (ERPs) that are used as the principle dependent variables throughout all
three studies.

1.1 The evidence that comprehension is predictive

As I outline below, many paradigms provide evidence that language comprehenders do, at least
in certain circumstances, make predictions about upcoming words while listening to an utterance or
reading a text. However, not all researchers agree on the extent to which comprehension is predictive,
and if it is, whether or not prediction should be treated as a fundamental linguistic process. For instance,
Jackendoff (2007) argues that it is unclear how prediction might directly affect processes involved in
generating meaning, as the statistics that support predictions are not directly reflective of the
representations (e.g. grammatical or semantic relationships) that are required for language
comprehension. This constitutes, in my opinion, an overly narrow point of view on which processes
should be considered important for comprehension – there are likely to be some general cognitive
mechanisms that are not directly related to the “core” functions of syntax and semantics, but nevertheless are indispensable for language competency. For example, by allowing the system to pre-emptively commit to part of an interpretation, prediction could free resources for other functions such as inferential reasoning or the planning of a response (during conversation, for instance). Furthermore, comprehension usually suffers when we have to strain to make out perceptual aspects of written or spoken words, suggesting that in easier contexts prediction may be facilitating our ability to fluently arrive at a meaning. Finally, predictions of form representations may reflect a deeper fact about cognition, and could provide critical evidence that high-level functions follow a more general predictive coding framework (see section 1.2.1 for more on this point). If predictions are a basic component of the broader perceptual system, then by inference they must play a role in any explanation of how the brain implements word recognition.

Below I review some of the most persuasive evidence that language comprehension is in fact predictive, paying particular attention to outstanding questions this dissertation is designed to address.

1.1.1 Prediction of lexical identity

A fundamental assumption made by the present work is that language representations are hierarchically structured (into word categories, words, phonemes, orthographic features, etc.) and that, in principle, predictions could be made at any of these levels. At the present time, however, evidence is strongest for prediction at the lexical level, i.e., at the (relatively) abstract level of a particular word and its meaning. Put another way, the most convincing evidence for prediction does not establish that predictions of upcoming words are also made at lower, perceptual levels of representation. Existing evidence for lower level prediction is more ambiguous, and will be discussed in the next section.

Some of the most-cited examples of word-level prediction come from studies of eye movements. Experiments using the “visual world” paradigm involve the tracking of eye movements
while participants listen to utterances describing an experimentally controlled visual scene. In one example, Altmann & Kamide (1999) tracked eye movements across a scene depicting a boy, a cake, a beach ball and toy train. When participants heard the sentence fragment “The boy is going to eat the …” their eyes tended to move more in anticipation to the picture of the cake, rather than the train or beach ball, prior to actually encountering the word “cake”. This was compared to eye movements across scenes containing more than one edible item, which did not show the same anticipatory pattern. Such results suggest that, at the very least, language comprehenders use what information happens to be available in to predict what future referents are likely to be mentioned.

To the extent that the results of Altmann & Kamide (1999) (and others using a similar paradigm, e.g., Kamide, Altmann & Haywood, 2003) can be taken as evidence for prediction, they indicate that it occurs at the most abstract, semantic levels of representation. For instance, while the eye movement pattern described above is compatible with predictions of particular words such as “cake”, “pastry”, “desert”, “treat”, “sweet” etc., it is also compatible with prediction of semantic features without commitment to a particular word or words (in this case, such features may be as abstract as “edible thing”). Indeed, an argument can be made that such abstract predictions could actually (on average) be more useful than predictions of specific words, as pre-activating a semantic feature may simultaneously facilitate the recognition of many potential words. Of course, more specific predictions may also confer the same benefit; “cake” also carries with it the feature of “edible thing,” and so the prediction of this particular word may equally facilitate the recognition of “pastry.” As alluded to earlier, it is an empirical issue as to whether the cost in prediction error (e.g., if “pastry” is encountered when “cake” was expected) would outweigh the facilitation due to their shared semantic features.

More constraining evidence for lexical-specific prediction comes from electrophysiological data. The N400 (Kutas & Hillyard, 1980), a negative going deflection in the language-related
Event Related Potential (ERP) around 400 msec. post-stimulus, has been found to be a reliable index of a word’s predictability, given its context. The basic result, now replicated hundreds of times (See Kutas & Federmeier, 2011 for a review), is that N400 amplitude increases as the likelihood of a word’s occurrence goes down, and is largest for words that are semantically anomalous. For example, the amplitude of the N400 in response to “socks” is greater than “butter” given the sentence context “He spread the warm toast with ...” (Kutas & Hillyard, 1980). However, while the correlation between N400 amplitude and predictability is well established, only a handful of studies provide direct evidence that the N400 reflects prediction per se. This is because in the standard experimental manipulation of semantic fit (e.g. the comparison of “socks” to “butter”), subsequent change in N400 amplitude can also be explained in non-predictive terms – for example, by positing a reactive process that evaluates the plausibility of a word in context after the word has been encountered.

Several studies looking at changes in the ERP occurring prior to an expected word provide more constraining evidence that the component’s underlying process may be predictive in nature. Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort (2005) report such a study using Dutch, a language in which adjectives must agree in grammatical gender with their co-occurring noun. The Authors constructed sentences in which the position of a predicted noun was preceded by an adjective that either matched or did not match the gender of the predicted word, allowing them to ascertain effects of prediction based on whether the gender of the adjective was surprising (given the expected noun), and thus before the expected word was encountered. For example, (English translation) “The burglar had no trouble locating the secret family safe. Of course, it was situated behind a [large neu painting neu] / [large com bookshelf com]”, where com indicates common gender, and neu indicates neuter gender. The authors found that adjectives whose gender was surprising given the expected noun

---

1 Original Dutch version of critical words, to illustrate the gender marking: groot neu schilderij neu / grote com boekenkast com (large painting / large bookcase).
("large_{com}\ldots bookshelf, in the previous example) elicited larger central-parietal positivities than adjectives with compatible genders. Because syntactic gender is an arbitrary feature of every word, and changes in the ERP recorded prior to the expected word ("painting") appeared to reflect this feature, the results indicate that participants were already predicting (at least) one particular word by the time the adjective (large_{neu}/ large_{com}) appeared in the input. A similar finding in English was made by Delong, Urbach & Kutas (2005), who found that N400s were reduced in response to a determiner that was phonologically congruent with a predictable but not-yet-presented noun (e.g., The day was breezy so the boy went outside to fly [a kite / an airplane]. Again, because the predicted word ("kite") is arbitrarily associated with a particular form of the determiner ("a"), a deviation in the brain response to "an" suggests participants had a specific expectation for how the sentence would continue (for still other similar results, see Wicha, Bates, Moreno & Kutas, 2003 and Wicha, Moreno & Kutas, 2004). Finally, we should note that a recent attempt to replicate the results of Delong, Urbach & Kutas (2005) failed (Ito, Martin & Nieuwland, 2016), suggesting that more work is needed to unambiguously establish the predictive nature of the N400.

The foregoing studies show that predictions can be specific to one particular word, rather than just a set of semantic features, since the features which elicited the effects (e.g., gender) are word-specific. Given the evidence that predictions can be lexically specific, it is natural to ask just how specific they can be – for instance, whether even low-level aspects of a word’s form might also be predicted during comprehension. As we will see in the next section, existing evidence for this is ambiguous, and thus constitutes the motivation for Study 1.

1.1.2 Prediction of word-forms

Some experimental evidence has been taken to suggest that language comprehenders make predictions about upcoming words at the level of perceptual features (e.g. visual form). Such
conclusions are usually based on the timing and scalp distribution of early EEG and MEG effects that are observed in response to unexpected sentence-embedded words. It is reasoned that if an effect is early enough (e.g., before 200 msec.), and largest over occipital cortex, then an early response to an unexpected word may suggest a mismatch between low-level visual predictions and bottom-up input.

Dambacher, Rolfs, Göllner, Kliegl & Jacobs (2009) made such an inference when they manipulated words embedded in German sentences that were either predictable or unpredictable given the context (e.g., (English translation) “The man on the picture fiddled around with models of Columbus’ fleet. In his right hand he held a [ ship / scepter ...”)). The authors found that the visual ERP response at occipital scalp channels was larger for unpredictable words as early as 90 msec. post-stimulus. These effects cannot be attributed to mere word-form familiarity because half of their stimulus sentences contained a high frequency target (“ship”) and the other half a low-frequency target (“scepter”, with a correspondingly appropriate context), and the effect of predictability did not interact with frequency. One question that immediately arises when considering these results is why most other studies using a similar manipulation of predictability have typically shown only N400 effects, but no early visual effects. The authors suggest that this may be due to a difference in stimulus onset asynchrony (SOA) – they used what they considered to be a more “natural” rate of stimulus presentation (280 msec. per word), which is shorter than many earlier studies that focused primarily on N400 effects. As a result, Dambacher et al. (2009) argue that their participants may have engaged in a more aggressively predictive strategy to help keep up with a more demanding pace. Finally, other studies have also reported rapid effects (within 150 ms.) of misspellings of predicted words (Kim & Lai 2012) and syntactic category violations (Dikker, Rabagliati, Farmer & Pylkänen, 2010). Incidentally, both of these studies used SOAs that were longer than Dambacher et al. (2009).

Even if the effects of Dambacher et al. (2009) and similar studies do reflect a rapid visual response to an unpredicted word, such an effect would still not constitute direct evidence for form
prediction. Similar to how most N400 effects of unexpected words can be accounted for without recourse to prediction, rapid effects with an apparent source in cortical areas may instead reflect a rapid detection of semantic incongruency. Although the latency of the N400 effect has led some to infer that access of semantic representations occurs for the first time at approximately 400 msec., there is now substantial evidence that such information can be retrieved much more rapidly, compatible with the possibility that early sensitivity to unexpected words need not reflect prediction. Indeed, several electrophysiological studies report effects of animacy (i.e., whether a word refers to something that can move on its own; Amsel, 2011) and other semantic features (Hauk, Davis, Ford, Pulvermüller & Marslen-Wilson, 2006) during single word recognition, within 200 msec. of stimulus onset. Because these effects occurred in the absence of any sentence context, and thus without affordances for prediction, they suggest rapid bottom-up access of semantic representations.

Thus while some studies provide evidence compatible with the prediction of word-form features, all such data (that I am aware of) can plausibly be accounted for by rapid bottom-up access of semantic features, which then conflict with representations of the prior context. Study 1 in this dissertation was designed to remove this ambiguity by showing effects of prediction when an expected stimulus is omitted from presentation. As described in more detail below, we presented participants with highly constraining sentences, one word at a time, and selectively withheld some critical, highly predictable words from presentation. We then asked whether brain responses at sentence positions with omitted words were correlated with the forms of the omitted words. Any such effects would then have to be due to prediction, since there is no bottom-up stimulus capable of generating conflict with contextual representations.

While we have seen that existing psycholinguistic evidence is ambiguous with respect to the possibility of word-form prediction, some more general models of perception imply that such
mechanisms should be consistently at play during language comprehension. We review some of these frameworks below with the aim of developing a more general-purpose language for discussing the mechanisms that may underlie any low-level predictions that may occur during language comprehension.

1.2.1 General Models of Prediction in Perception and Cognition

While there is increasing evidence for prediction during normal language comprehension, its relationship to specific language processes remains ambiguous. Some have argued that prediction should be treated as one of several strategies available to language comprehenders, and that as such it reflects an opportunistic process that occurs only when there are resources to support its use (Federmeier, 2007). In other architectures, such as early connectionist models of language processing (e.g. Elman, 1990), prediction occurs as a fundamental consequence of how the models are trained to discover the statistical dependencies between elements of a sequence. To learn the dependencies between consecutive words, for example, recurrent networks are trained to produce (i.e., predict) the next word in a sequence, given the current word and some representation of the model’s previous state. For these models, prediction is computationally fundamental, rather than a high level strategy that can be turned on or off.

Although recurrent connectionist models of language processing are not always intended to be interpreted as models of brain function, other work has more directly made the claim that prediction represents a fundamental type of neural computation. In this section I outline some of these “predictive coding” models and what their implications are likely to be for the processes underlying word recognition. We will see that they suggest (1) that form-level predictions should be computed as a matter of course, and (2) that the computation of predictions and prediction error (e.g. the surprise due
to an unexpected word-form) should be segregated as separable neural processes. Finally, we will see how these two implications inform the design of studies 2 and 3.

It has been posited that prediction and prediction error are the fundamental pieces of information that are passed between hierarchically organized cortical areas (e.g. through the top-down connections between areas V2 and V1 in the visual cortex; Rao & Ballard, 1997; Mumford, 1992). In particular, Rao & Ballard (1999) model the hierarchy of the visual system as a series of layers containing units of increasing receptive field size. Central to this organization is that higher levels generate predictions about the activities of the level immediately below, and lower levels transmit the difference between these predictions and their current estimate of the input (the prediction error) back up the hierarchy as feed-forward information. Prediction errors are then used to adjust the model so it may more accurately represent the current stimulus, as well as any future stimuli with a similar pattern. Critically, Rao & Ballard (1999) implement this architecture by positing two separate types of neural units: representation units that maintain the layer’s current estimate of the input, and provide the prediction to the level immediately below; and error units that compute the difference between the activity of the layer’s representation units and the top-down prediction received from the layer immediately above. Interestingly, findings from cat (Murphy & Sillito 1987) and monkey (Zipser, Lamme & Schiller, 1996) physiology suggest that the response profiles of neurons in layers 2 and 3 of the visual cortex may be good candidates for serving as analogs of Rao & Ballard (1999)’s error-detecting units.

Several researchers have claimed that the core aspects of Rao & Ballard’s (1999) model could represent a more general architecture for the brain’s perceptual and cognitive systems (Clark, 2013; Huang & Rao, 2011; Friston & Seibel, 2009; Lee & Mumford, 2003). In particular, these “predictive coding” frameworks represent a revision of the traditional view of stimulus recognition as implemented by a relatively static hierarchy of feature detectors, whose individual units are tuned to a relatively
unchanging set of stimulus features and thus serve primarily as bottom-up stimulus decoders (Riesenhuber & Poggio, 2000; Hubel & Wiesel, 1965). In predictive coding models, the brain’s perceptual channels are instead modeled as a duplex system (Clark, 2013) that maintains two complementary types of representations: stimulus expectations (i.e. predictions) and prediction error (similar to the two distinct types of neural units by Rao & Ballard, 1999). More specifically, probability distributions over stimulus expectations are encoded at each level of the perceptual hierarchy and transmitted to lower levels, while expectations at the current level are compared to incoming top-down predictions and forwarded as prediction error to higher levels. Among other things, this means that brain activity observed in response to a stimulus is likely to represent a summation of both prediction generation and prediction error, which may vary independently. More concretely, if we record from a part of the brain that we believe to process word-forms, then the observed neural activity in response to a word will reflect both the expectation for that word and the surprise associated with it.

As an example, consider how a predictive coding architecture would explain the responses of the brain’s visual word form area (VWFA, a part of visual cortex that has been argued to be selective for orthography; McCandliss, Cohen & Dehaene, 2003) to either a visual word or (for example) a picture of a house. Under the classical view of perception as bottom-up feature detection, the VWFA should respond stronger to words than houses, regardless of expectation level (perhaps the baseline level of response varies with other factors such as attention, but the relative difference in response to words and houses should remain constant). Under a predictive coding architecture, however, the VWFA will differentially encode information related to the expectation of, and surprise (prediction error) to, a particular word, and that these two factors will combine to determine the measured brain signal. Assume first that expectation for a particular word is high. First, this means that the activity of the VWFA related to expectation will be high, regardless of whether a word or a house was encountered. Under high expectation, if a word is encountered, then units representing prediction error will contribute only a
small additional amount to the brain signal, and so the response difference between houses and words will be relatively small. Now assume that expectation for a word is low – in this case the VWFA’s response to a word will be dominated by prediction error, increasing the difference between the response to a word and the response to a house (the VWFA’s response to the house will only reflect the expectation for a word, which is low). Thus, somewhat counterintuitively, the predictive coding architecture expects a larger difference in the responses to words and houses during lower levels of expectation for words, since there will be higher levels of surprise in response to words than to houses. Egner, Monti & Summerfield (2010) ran exactly this experiment, only using faces and houses as stimuli, and recording from Fusiform Face Area (FFA, a region of visual cortex thought to be relatively selective for faces). As predicted, they found that the difference in the FFA’s response to the two stimuli was largest when expectation for a face was lowest.

Thus, if it is indeed the case that prediction and prediction error constitute a basic division of labor within the word recognition system (and the brain more generally) it suggests that the expectation for a word and the prediction error in response to a word should be differentiable in the brain signal. Furthermore, because the result of any error computation mechanism inherently depends on the presence of some kind of prediction, the temporal sequencing of the two operations may have a clear ordering: expectation generation first, then error computation. We test this hypothesis in study 2, reported below.

---

2 Here we must be careful when discussing the relative timing of expectation/prediction generation and the processing of prediction error, given the well-established presence of cortical loops in which two areas can be mutually self-reinforcing (or inhibiting). The presence of such loops certainly complicates the establishment of “primary cause” in an information processing system. However, given the millisecond-level temporal resolution of measures like EEG, combined with the requirement that prediction generation logically must precede detection of prediction error, it may be that at early visual stages of word recognition (where there simply hasn’t been enough time for activation to ‘reverberate’ between multiple cortical levels) may nevertheless be a good candidate for a stage of processing in which the two processes can be clearly isolated in time.
In order to be most clear about how top-down mechanisms of prediction generation and the bottom-up mechanisms of prediction error should be mapped to the process of word-form recognition, it is necessary to describe a general framework for how predictions at more abstract levels of representation (e.g. semantic or lexical) may in turn drive predictions at lower form-based levels. First, such a claim makes the assumption that language representations are in fact hierarchically structured, and that levels of the hierarchy are bi-directionally connected. Given the substantial evidence not only for the hierarchical nature of language in the abstract (going back to Chomsky, 1957), but now for the hierarchical nature of the ventral visual object recognition stream in general (Felleman & Van Essen, 1991), such an assumption appears warranted. To be sure, there is more disagreement about the precise nature of the representations at each level of this hierarchy, and (for instance) the extent to which they map on to the division of syntax, semantics, morphology, and phonology/orthography, as assumed by dominant (but sometimes biologically ambivalent) models of language. For our purposes, we make the minimal assumption that predictions are initially generated at higher levels of word representation that encode lexical identity and semantic content.

Thus if “predictive coding” is taken to be an underlying principle of perception, we should expect that top-down predictions of upcoming word forms should happen as its natural consequence. It is important to be clear, however, that predictive coding by itself does not directly explain how the initial predictions about upcoming words are generated, e.g. how one comes to expect “sugar” given the words “I take my coffee with cream and ...”. These high level predictions may arise via multiple mechanisms, e.g., by lateral spreading of activation in some cases and by top-down retrieval in others. As reviewed in the previous section, there is substantial independent evidence that such predictions do in fact occur, but only in combination with a predictive coding mechanism would we naturally expect additional low-level anticipation to occur. In other words, given the pre-activation of a word at a higher
level of representation, predictions about the activity at lower levels should automatically be generated. Such predictions should occur whenever representations at higher levels are active.

Top-down predictions of word-forms may be a direct consequence of a predictive coding architecture, but we must also consider the feedforward passing of prediction error, which constitutes the other core component of the duplex models of Rao & Ballard (1999) and others. Again, in these models each layer of the processing hierarchy maintains two separate types of units: one group dedicated to representing its current estimate of the input (which gets back-propagated to as the level immediately below as predictions) and another dedicated to computing the discrepancy between the current estimate of the input and the predictions received from higher levels (which serves as the feedforward signal). In terms of word-form recognition, this suggests that activity at lower levels (such as those representing form features) which do not match the predictions imposed by higher levels (such as those representing lexico-semantics) should be represented as a brain response that is proportional in size to the discrepancy. In most language processing contexts this discrepancy will be correlated with the extent to which the stimulus was expected. While this seems like a natural account of the brain’s response to a linguistic stimulus, we should keep in mind that it is not the only possible such driver. Some accounts, for instance, emphasize the extent to which an incoming word fits with the previous context. These accounts might be termed “predictive” in a passive sense, as information activated by the prior linguistic context sometimes shares features with the incoming word, thereby facilitating its identification. However, this notion of prediction is distinct from the sense of prediction that we have been using thus far – the idea that upcoming words and their forms are pre-activated prior to their arrival in the input (see section 5.1 as well as Van Petten & Luka, 2012 and Kuperberg & Jaeger, 2016 for more on the nuances of prediction vs. pre-activation). Thus, the brain response to an unexpected word may be explained by the extent to which it deviates from prior input, without positing a more actively predictive system (see sections 1.1.1 and 1.1.2). Study 1 aims to provide more direct evidence for pre-
activation specifically, by omitting highly predictable words from a sentence, and observing whether or not the resulting brain response still reflects arbitrary features of their form.

1.3 Early Brain Responses to Visual Words and their Potential Link to Prediction

The foregoing discussion has outlined predictive coding as a potentially general model of perception. Central to this framework is prediction-generation and error-computation as two mutually dependent mechanisms that (in at least some versions) are implemented in distinct neural circuits. Considering well-known neural indices of word-recognition, is it possible to identify potential candidates for word-form prediction and computation of prediction error? One such candidate, the P1 ERP component, may at least partially fit the bill, though it has not previously been interpreted in such a way.

The P1 ERP and A Potential Link to Prediction

The visual P1 ERP component (sometimes called the P100) peaks between 90-130 msec. after the onset of a visual stimulus, with its largest amplitudes typically recorded from occipital scalp channels that sit over visual cortex. Most research has assumed that the component is “exogenous”, in the sense that its amplitude seems to be closely related to low-level visual characteristics of the stimulus, e.g. word length (Hauk, Davis, Ford, Pulvermüller & Marslen-Wilson, 2006), basic facial features (Hermann, Ehlis, Muelberger & Fallgatter, 2005) and color information (Zhang & Luck, 2008).

Perhaps the most documented feature of the P1 is its sensitivity to spatial attention. For example, it has been shown that selective attention to one part of space increases the amplitude of the P1 component when a stimulus appears in the attended (as opposed to unattended) region of space (see Hillyard & Anllo-Vento, 1998 for a review). This has been interpreted as reflecting a gain control function that increases the signal-to-noise ratio of neurons whose receptive fields overlap with the
attended region. These results are also consistent with so-called “early selection” models of attention, in which low-level sensory inputs are modulated before the stimulus can be categorized (P1 amplitudes tend to be similarly enhanced to all objects appearing in an attended space, whether or not they are “targets”; Posner & Dehaene, 1994).

I argue that an alternative way to frame effects of selective attention on the P1 is to say that the attended-to spatial location reflects the participant’s “prediction” about where the stimulus will appear. In these terms, enhancement of the P1 during validly-cued trials could instead reflect the result of top-down predictions about the location of the stimulus that are imposed by higher cortical areas, with the representational units responsible for that region of space tuned in anticipation. Thus, in predictive terms, allocation of exogenous spatial attention is equivalent to anticipatory tuning of the cells in visual cortex that will be recruited to handle the computations needed to process the upcoming stimulus. Indeed, we can use a similar logic to that of Egner, Monti & Summerfield (2010) to explain P1 attention effects. If the response to a stimulus can be decomposed into prediction and prediction error, then we can re-cast anticipatory attention to one part of space as near-certain expectation, which is revealed as an increase in P1. On the other hand, a response to the stimulus presented to the unattended side will be dominated instead by prediction error, presumably in proportion to its probability, but will not reflect much expectation. Thus, in a context where spatial cues are 70% valid, response to a validly cued stimulus will reflect something close to 100% expectation in addition to 30% prediction error, while response to an invalidly cued stimulus will be proportional to something like 0% expectation in addition to 70% prediction error. This correctly predicts the observed asymmetry in P1 amplitude between validly and invalidly cued responses.

This way of thinking gives us the tools to make concrete predictions about how P1 amplitude should be modulated by both top-down expectations and surprise during reading. First, prediction of an
upcoming word should increase expectation in cells that are retinotopically mapped to the areas of space that will overlap with the word-form. Assuming the traditional retinotopic mapping of visual cortex, expectation for a longer word should then modulate the activity of a larger number of cortical units. Within each patch of visual space, we then predict that separate units will encode the expectation for a particular feature, as well as the prediction error due to the feature that was actually encountered. Importantly, expectation in this case may not be totally conflated with spatial attention (as in the Posner cuing task), but rather proportional to the confidence with which the form-feature is predicted. Surprise, on the other hand, should be largest in areas outside of the expected spatial extent of the word, where no strong top-down expectations were received.

This line of reasoning leads to some hypotheses about how the P1 should be modulated during reading. In particular, P1 amplitude should reflect a mixture of prediction and prediction error, in proportion to the certainty with which a particular word-form is expected. To give an example, suppose that a reader has an expectation for a particular four letter word. In the predictive coding framework outlined above, this would imply a top-down modulation of cells with receptive fields overlapping with expected horizontal extent of the word, and this should be reflected in a P1 amplitude that is proportional to the expected word’s length. Second, if the reader encounters a word that differs from their expectations, then the aspect of the brain response encoding surprise will be maximal outside the visual space that the expected word was to occupy. Inside the visual space of the expected word, surprise will be proportional to the extent to which the input features happen to match those of the expected word.

The practical consequence of these predictions is that P1 amplitude should be modulated both in relation to an expected word’s length, as well as a function of how “off” or surprising the actual input word is (i.e., the prediction error).
1.4 The current studies.

The studies described here aim to rigorously test whether the language system implements top-down predictions about the form features of upcoming words. This general question is approached by testing four different sub-hypotheses, referred to in the rest of the dissertation as SH1-4. The first is that expectations for particular word forms can be observed independently of the form of the actual sensory input. Second, if the brain’s early stages of word recognition are served by a predictive coding mechanism, then we should be able to observe independent effects of prediction and prediction error. Third, we should also be able to find that any effects of prediction should precede effects of prediction error, as is logically required by a predictive coding architecture. Fourth and finally, if prediction is essential to word recognition in general, then we should observe it in contexts outside of those that strongly afford for the prediction of one particular word.

Each of the three sub-hypotheses addresses important theoretical issues surrounding the brain’s architecture for comprehension. SH1 aims to reduce the ambiguity of previous results and firmly establish the existence of top-down predictions about the perceptual features of upcoming language. SH2-3 look to establish a clear separation and temporal sequencing of prediction generation and error computation, which are the two fundamental quantities of information in predictive coding architectures. SH4 seeks to establish whether prediction of perceptual features is a consistent feature of reading, a result that would suggest that the basic neural infrastructure for word recognition is fundamentally predictive.

Finally, this research can be viewed as a broader theoretical attempt to bridge one of the few general-purpose accounts of brain function, and the specific (but complex) case of word recognition during language comprehension. As noted above, predictive coding models of perception represent a general claim about the nature of the messages that are passed between hierarchies of the brain. While
these frameworks are often highly abstract, and meant to be general accounts of neural computation, there has not been much evidence that they can account for a complex function like word recognition in context. Thus the research presented here can be read from two different points of view: as an empirical investigation of word-form prediction during comprehension, and as a hard test-case for predictive coding.

2. Study 1: Early EEG Responses Betray the Forms of Words that Were Predicted but Never Seen

As mentioned in section 1.1.2, some evidence is compatible with the pre-activation of perceptual features during comprehension (Dambacher et al., 2009, Dikker et al., 2010, Kim & Lai, 2012), but the speed with which semantic representations can be accessed from the bottom-up means that such early effects may in fact reflect semantic mis-match effects rather than true form pre-activation. In Study 1, we used a new paradigm that provided a more constraining test of whether such form-feature pre-activation in fact occurs (SH1). We recorded EEG from the scalp while participants read sentences that each contained a predictable word. In half of these sentences, we omitted the predictable word from presentation, replacing it with a simple vertical crosshair stimulus (Figure 1.1). We then asked whether brain responses to these vertical crosshairs were sensitive to the form features of the omitted word by focusing on the relationship between the occipital-temporal P1 ERP and the predicted word’s character length.

Prior studies have established the relationship between early events in the occipital EEG and the character length of visually presented words (Hauk et al., 2006). Furthermore, fMRI-based research has supported the conclusion that higher order cortex imposes top-down modulation of sensory cortices (e.g., Kok, Jehee, & de Lange, 2012; Summerfield & Koechlin, 2008). However, it is still unclear whether such top-down interactions actually represent predictive processes (given the poor temporal resolution of fMRI; Hauk, 2016). In the current study, we reasoned that if a relationship between brain activity and
word length can be observed when words are omitted from the visual input, then such a correlation 
would have to be interpreted as reflecting prediction, rather than bottom-up analysis or late-stage top-
down interactions.

2.1 Methods

2.1.1 Participants

Participants were 34 right-handed native English speakers from the University of Colorado community (ages 18-25, mean 19.7; 13 female) with normal or corrected-to-normal vision. One 
participant’s data was excluded due to data recording malfunction, and seven were excluded due to less 
than 75% of their data being retained after EEG pre-processing and/or removal of incorrect responses 
during the sentence processing task (see below). Data from 26 participants were analyzed. Participants 
were compensated with either 20 USD or course credit. For this and all other studies reported in this 
manuscript, participants gave their written, informed consent to participate, and all study procedures 
were approved by the University of Colorado IRB prior to data collection (IRB protocol number 15-0802).

2.1.2 Stimuli

200 sentences were created, each containing one critical noun that was highly predictable given 
its sentence context (see Table 1 for examples). Half of the critical nouns were short (range 3-6 
characters; mean = 4.5) and half were long (range 7-14 characters; mean = 8.9). The length of each 
sentence context prior to the critical word varied between 4 and 20 words, and the average length of 
contexts preceding short and long words was matched (12.49 words before critical short words and 
12.11 words before critical long words; two-sample t-test of the differences t(198) = .97, p = .39). Critical 
nouns were never in sentence-final position, to avoid contamination from end-of-sentence wrap-up
effects (Hagoort, 2003). Each sentence could appear in a form that omitted the critical word, leaving a crosshair in its place (omitted word condition) or with the critical word unaltered (word present condition; see Figure 1.1 for an illustration). Two presentation lists were created, each containing one occurrence of each of the 200 sentence stimuli. Within each presentation list, half of the long critical words and half of the short critical words were omitted from their sentences, and the other half of the critical words were present in their sentences. Any sentence that occurred in the omitted word condition within one list occurred in the word present condition within the other list. Thus, each presentation list contained 50 sentences from each of four within-subjects experimental conditions (short word presented, short word omitted, long word presented, long word omitted). Sentences were pseudo-randomly ordered within each list with the constraint that no more than three sentences from the same condition occurred sequentially. Each participant saw one list.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Pre-target Sentence Context</th>
<th>Critical Word</th>
<th>Post-target Sentence Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>They could not make up their minds so they decided by flipping a coin they found on the ground.</td>
<td>coin</td>
<td>they found on the ground.</td>
</tr>
<tr>
<td>Short</td>
<td>Vicki took in all and dove underwater to see how long she could hold her breath this time.</td>
<td>breath</td>
<td>this time.</td>
</tr>
<tr>
<td>Short</td>
<td>When Vincent abandoned his friends and ideals for a big contract, people said that he had sold his soul to the devil.</td>
<td>soul</td>
<td>to the devil.</td>
</tr>
<tr>
<td>Long</td>
<td>The young parents bought their son a big cake and lots of toys to celebrate his birthday last week.</td>
<td>birthday</td>
<td>last week.</td>
</tr>
<tr>
<td>Long</td>
<td>Sergio had been a successful professional chef for years when he finally decided to open a restaurant for himself.</td>
<td>restaurant</td>
<td>for himself.</td>
</tr>
<tr>
<td>Long</td>
<td>The football player was taken out because he was hit in the head and suffered a concussion from a tackle.</td>
<td>concussion</td>
<td>from a tackle.</td>
</tr>
<tr>
<td>Long</td>
<td>My parents got married 50 years ago today so they are going to celebrate their anniversary on a cruise.</td>
<td>anniversary</td>
<td>on a cruise.</td>
</tr>
</tbody>
</table>

**Table 1: Example stimulus sentences for study 1.**

Semantic support for the critical words was quantified by asking participants on Amazon Mechanical Turk to read the experimental sentences up to but excluding the critical word, and to produce the “best completion” (cloze ratings). Critical words used in the experiment were produced in at least 85% of responses to their sentence context (mean cloze short word condition = 94.9%, long word condition = 95.3%, two-sample t-test t(198) = .50, p = .62). At least 20 different MTurk participants provided completions for each sentence stimulus.
Finally, although long and short critical words were chosen to be matched on their frequency in the language as closely as possible, there still remained a significant difference between the average natural log of their frequencies, as estimated from the CELEX corpus (Baayen, Piepenbrock & Gulikers, 1995; two-tailed, two-sample t-test: t(198) = 3.00, p = .003; short word mean log frequency = 4.27, long word mean log frequency = 4.03). As such, we included Log Word Frequency as a covariate in all of our regression analyses (see below).

2.1.3 Procedure

EEG recording took place in a dimly lit, sound-shielded booth with subjects seated 140 cm in front of an LCD computer monitor. Sentence stimuli were presented one word at a time in white fixed-width font (Courier New) in the center of a black screen (Rapid Serial Visual Presentation). Each character occupied ~.8 degrees of visual angle. Each word was on the screen for 300 msec. followed by a 300 msec. blank screen (stimulus onset asynchrony = 600 msec.). Each word was accompanied on the display screen by a vertical crosshair above and below its mid-point (Figure 1.1). In cases where the critical word was omitted only the vertical crosshair was displayed. After each sentence, a screen appeared asking participants whether they noticed an omitted word. If they answered yes, they were asked to name the word by speaking into a microphone. Answers were coded online by experimenters. Participants saw 6 example stimuli (3 sentences with the critical word omitted and 3 sentences with the critical word presented) prior to starting the experiment. An experimenter remained in the room with the participant to answer any questions during the practice period. Sentence stimuli were presented in four blocks of 50 sentences each, with a break between blocks.
Figure 1.1 Stimulus sentences were displayed one word at a time in Rapid Serial Visual Presentation format. Each word was displayed for 300 msec. within a vertical crosshair, followed by a blank screen lasting 300 msec. (not shown). Sentence contexts were predictive of either a short word (left) or long word (right). In one half of the sentences, the target word was displayed, and in the other half, it was omitted, leaving only the vertical crosshair (bottom).

2.1.4 Analysis of Behavioral Data.

Participant’s mean accuracy in identifying sentences with an omitted word was 98.9% (mean 98.9%; range 88%-100%). When participants confirmed that a word was omitted, they correctly named that omitted word in 89% of cases (range 72%-97%). Only trials in which participants correctly named an omitted word were included in EEG analysis (94.5% of all trials prior to EEG artifact rejection).

2.1.5 EEG Data Pre-processing.
Continuous EEG was recorded from 66 Ag/AgCl electrodes (Neuroscan Quik-Caps with SynAmps 2 amplifier) at 1000 Hz referenced to a vertex electrode. Subsequent to acquisition, data were down-sampled to 200 Hz, and band-pass filtered between .1 and 50 Hz. Epochs of data spanning -200 msec. to +1000 msec. relative to word onset were extracted for all words in the stimulus sentences, and baseline corrected to the 200 msec. pre-stimulus time interval. Epochs containing voltage movements exceeding +/- 100 microvolts were rejected as artifactual. All EEG data preprocessing was done with the Neuroscan Edit software.

2.1.6 Analysis of P1 Event Related Potential

We analyzed the effect Word Length on the P1 ERP. For ERP analyses, EEG data were re-referenced to the average of all scalp electrodes. Each participant’s epoched dataset was normalized at each time-sample using each participant’s mean and standard deviation across all epochs for that sample (z-score transform). This was done to minimize the effect of any individual differences in ERP component size on our analyses designed to predict P1 amplitude from character length. We then pursued two distinct but related types of analyses. First, we examined the effect of word length on the P1 response to all non-critical target words (i.e. all surrounding context words), in order to ascertain the “typical” effect of word length on P1 amplitude. Second, we examined the effect of word length on the P1 response to all critical target words, including any potential interactions with our other independent variable of interest, Word Presence. This was done to test our hypothesis that P1 amplitude will be related to the lengths of our critical target words, even when the critical word was omitted from display.

For the target words, each subject’s data was averaged within each of the four conditions (short-presented, long-presented, short-omitted, long-omitted). ERP averaging was done for display purposes only (see Figure 1.2); all regression analyses reported below treat word length as a continuous predictor.
Figure 2.2: Relationship between Character Length of all non-target words and P1 amplitude at Left, Center and Right Posterior scalp locations (left to right). X-axes are word length, Y-axes are normalized P1 Amplitude. The radii of the red circles are proportional to the number of observations at that Character Length. Small black circles are the average component activations for each participant at each Character Length. Bands about the regression lines are 95% confidence intervals. Scalp montages above each plot indicate the electrodes used in statistical analyses for that location.

Left: There was only a marginally significant relationship between the lengths of non-target words and the amplitude of the P1 at Left Posterior electrodes. Left and Right: Both Center and Right electrode groups showed a significant positive correlation between non-target word length and P1 amplitude.

We quantified ERPs in a time window spanning 120-160 msec. post-stimulus onset, intended to capture the peak of the P1. This time-window was chosen based on previous studies in our lab which have consistently found P1 peaks around 135 msec. post-stimulus (e.g. Kim & Lai, 2012). Within this time window, ERPs were calculated at three different groups of posterior scalp electrodes, chosen for their position over visual cortex: P3, P5, P7, PO7, PO5 (left posterior electrode group); OZ, O1, O2, CB1, CB2 (midline posterior electrode group); P4, P6, P8, PO9, PO6 (right posterior electrode group).
P1 amplitudes were analyzed with mixed effects regression models including the following fixed factors: Left-to-Right Scalp Location (Left = -1, Center = 0, Right = +1), Lateral-to-Midline Scalp Location (Left = -1, Right = -1, Center = +2), Word Length (number of characters), and Log Word Frequency as estimated from the CELEX corpus. For analyses of critical target words, we also included the additional factor of Target Word Presence (+1 for presented words and -1 for omitted words). Interactions between Word Length, Scalp Location, and Word Presence were also included (where appropriate). Random intercepts for participant and stimulus item were also included, as well as random slopes for Word Length for each participant. All models were fit using the lme4 package for R (Bates, Maechler, Bolker & Walker, 2015). P-values for predictor variables were obtained from the lmerTest R package (Kuznetsova, Brockhoff & Christensen, 2016). Analysis of non-critical target words encompassed 68421 total observations, while analysis of critical target words included 2395 observations.

2.1.7 Independent Components Analysis

In order to separate different aspects of the brain responses to our stimuli, we used Independent Components Analysis (ICA) for blind source separation of our data, as there are likely to be several factors that influence the visual EEG response, and that these would be mixed together at similar areas of the scalp. Furthermore, our design involved two different types of stimuli (presented and omitted words) that may have had overlapping but nevertheless distinct response profiles. Transforming our data into statistically independent components was thus an effective way to identify distinct aspects of the EEG signal that were related to each type of stimulus.

In general, the ICA solution re-represents the EEG signal as a series of statistically independent components, whose activations at each point in time indicate their relative contribution to the original data. The activations of the ICs can be analyzed for experimental effects in a similar way to traditional ERPs (e.g. Kim & Gilley, 2013, Makeig et al., 2002).
All ICA analysis was conducted using the BINICA infomax algorithm within the EEGLAB package for Matlab (Delorme & Makeig, 2004). ICA was applied to a group-level dataset constructed by horizontally concatenating the pre-processed, vertex-referenced and epoched EEG data from all participants. The input dataset included epochs of data that spanned the entire sentence stimulus (i.e. not just responses to the critical words). We included data from all electrodes except for the ocular channels (VEO and HEO).

2.2 Results

2.2.1 P1 ERP Results

Results of statistical analyses of the scalp ERPs are summarized in Table 3 and described below.

Analyses of all Non-Critical Context Words

Across all non-critical context words, we observed that longer words tended to evoke larger P1 amplitudes, as evident in a main effect of character length collapsing across scalp location ($\beta = .009, t = 4.90, p < .001$). P1 amplitude in general was found to be larger in right than left scalp sites, as reflected in a main effect of Left-to-Right Scalp location ($\beta = .026, t = 4.49, p < .001$; Figure 1.2). Furthermore, we found that the effect of Word Length on P1 amplitude also depended on Scalp Location, as reflected in an interaction between Word Length and Left-to-Right Scalp Location ($\beta = -.006, t = -5.00, p < .001$). To determine exactly where on the scalp the effect of word length appeared strongest, we followed this interaction with analyses of Word Length at each of the three scalp locations separately. This showed that the relationship between Word Length and P1 amplitude was only marginally significant at left posterior scalp channels ($\beta = .004, t = 1.72, p = .086$; Figure 1.2, Left), but more reliable and Center and
Right posterior channels (Center: $\beta = .010, t = 4.39, p < .001$; Right: $\beta = .013, t = 5.70, p < .001$; Figure 1.2 Center and Right).

Analyses of Critical Target Words

In the analysis of our critical sentence positions we again found a positive relationship between character length and P1 amplitude (main effect of Character Length; $\beta = .014, t = 2.13, p = .034$), as well as an indication that presented words evoked larger P1 amplitudes than omitted words (main effect of Word Presence; $\beta = .272, t = 5.67, p < .001$). Furthermore, the length of omitted words correlated positively with the amplitude of P1 at left hemisphere electrodes, while the lengths of presented words correlated positively with the P1 amplitude at right hemisphere electrodes (Figure 1.3). This was supported by a three-way interaction between Word Length, Left-to-Right Scalp Location and Word Presence ($\beta = .022, t = 2.64, p = .008$). We followed up this interaction with separate analyses at each electrode group that included the factors of Word Length and Word Presence.

At the left posterior electrode group, we found that increasing character length led to larger P1 amplitudes, reflected in a main effect of Word Length ($\beta = .020, t = 2.15, p = .034$). We also found evidence that this relationship depended on whether the word was presented, such that the effect only held for omitted words (Figure 1.3). This was supported by an interaction between Word Length and Word Presence ($\beta = -.032, t = -2.78, p = .005$). Considering only omitted words, P1 amplitude increased with word length ($\beta = .019, t = 2.14, p = .033$), but considering only presented words, the effect of length did not reach significance ($p = .21$).

At the right posterior electrode group, we observed a main effect of word presence ($\beta = .216, t = 2.59, p = .010$) but not of Word Length ($p = .25$) nor was there an interaction between Word Length and
Word Presence ($p = .37$). Given the three-way interaction between Word Presence, Word Length and Left-to-Right Scalp Location mentioned above, as well as the reliable correlation between Word Length and P1 amplitude found for all non-critical words, we explored whether right posterior P1 amplitudes were differentially related to the lengths of presented and omitted words. We found that P1 amplitude had a marginally positive relationship with the lengths of presented words ($\beta = .018$, $t = 1.94$, $p = .054$; Figure 1.3b), but not with the lengths of omitted words ($p > .15$).

At the midline posterior electrode group, presented words lead to larger P1’s than omitted words ($\beta = .22$, $t = 2.63$, $p = .008$), but we did not observe any effects of word length or any interactions ($p’s > .30$).
Figure 1.3 Scalp P1 ERP effects of target word length. Left panel: omitted words. Right panel: present words. The topographic plots (upper left corner of each panel) reflect the differences in voltage 120-160 msec. post-stimulus between long and short words, with red colors indicating more positive voltages for longer words and blue colors more positive voltages for shorter words. The front of the scalp is at the top of all topographic plots; left is left and right is right. The ERP traces (upper right corner of each panel) are plotted negative up, with a y-axis scale ranging +/- 3 microvolts. The x-axes of the ERP traces are time, ranging from -200 to 300 msec., with 0 representing the time of word onset. Red traces reflect the response to long words (7-14 characters long) and black traces the responses to short words (3-6 characters). Polygons overlaid on the inset electrode montage outline the electrodes used to generate the ERP trace. X-axes for scatter plots (bottom of each panel) are Character Length. Y-axes are normalized P1 amplitude 120-160 msec. post-stimulus from the electrode group used to generate the ERP trace. The radii of the red circles are proportional to the number of observations at that Character Length. Small black circles are the average P1 amplitude for each participant at each Character Length. A) For omitted words, longer Character Lengths elicited larger P1 amplitudes than shorter ones at left posterior electrodes. This can be seen in the red at left posterior sites in the topographic plot, the difference between the P1 amplitudes at 130 msec. in the ERP traces, and the positive correlation shown in the scatterplot. B) For present words, longer Character Lengths elicited larger P1 amplitudes than shorter ones at right posterior electrodes. This can be seen in the red at right posterior sites in the topographic plot, the difference between the P1 amplitudes at 130 msec. in the ERP traces, and the positive correlation shown in the scatterplot.
2.2.2 ICA Results

Our ICA returned 64 independent components (IC’s) for analysis. In order to focus on brain activity that reflected early responses in visual cortex, we elected to analyze only IC’s that met three criteria, which we developed as an objective, unbiased way to identify components that are likely to reflect the brain’s early response to a visual word. We decided that the component must (1) be one of the 10 components that explained the most variance in the original EEG data; (2) contain a peak between 100 and 200 msec. after the onset of the word, whose amplitude is larger than 1 standard deviation of the component’s activation throughout the epoch; and (3) have a scalp projection with a maximum that is posterior to the electrode CPZ (The scalp distribution of each component is plotted by mapping the signal mixing matrix onto the coordinates of the electrode montage).

Four independent components met these criteria and are referred to below as Components 1-4. Component 1 accounted for 6.4% of variance (3rd overall), had an activation peak at 140 msec., and a maximum scalp projection at electrode P5 (left posterior). Component 2 accounted for 4.0% of variance (6th overall), had an activation peak at 140 msec., and a maximum scalp projection at electrode CB1 (center-left posterior). Component 3 accounted for 2.6% of variance (7th overall), had an activation peak at 150 msec., and a maximum scalp projection at electrode O2 (center-right posterior). Finally, component 10 accounted for 2.0% of variance (10th overall), had an activation peak at 130 msec., and a maximum scalp projection at electrode PO6 (right posterior). The top two components in PVAF were not analyzed because the first lacked a clear maximum in its scalp projection and the second clearly reflected activity related to eye blinks. The scalp projections of components ranking 4th and 5th were too far anterior and the peaks of activation for components ranking 8th and 9th occurring before 200
msec. did not meet our threshold for amplitude. See Figure 1.4 for plots of the components selected for analysis.

Figure 1.4 Component activation time series and scalp distributions of (a) the four independent components selected for statistical analysis (see Methods section for component selection criteria) and (b) the remaining six, unanalyzed ICs ranking in the top 10 in PVAF. For each component, illustrated are: 1) PVAF (percentage on top of scalp projection), 2) scalp topography of the IC’s activity, 3) Component event-related time series computed by averaging the activation of each component in 1000 msec. epochs of time after the onset of every word (both target and non-target words). Component activation is plotted on y-axes of component time series. As with all ICA solutions, the polarity of the component time series, the units of component activation, and the units of the scalp projections are arbitrary.

Our statistical analyses of these four independent components mirrored those of our ERP effects. In particular, we started by analyzing the relationship between Independent Component activation and the lengths of all non-critical context words in analyses that included only Word Length
and Log Word Frequency. We then moved to analysis of our critical target words with models that included Word Length, Word Presence, Log Word Frequency, and the interaction between Word Length and Word Presence. All random effects structures mirrored those of the ERP analyses. For our dependent measure we quantified the average activation of each component in the time-window of 120-160 msec. for each epoch of data associated with the brain’s response to a word. Because we used the same test for each of the four components, we only considered a relationship between IC activation and our two predictors statistically significant if the resulting p-value was less than the Bonferroni-corrected critical value of .0125.

Below we describe the essential findings.

Independent Component 1

In our analysis of all non-critical words, we found a marginally significant negative relationship between the activation of component 1 (IC1, 6.4 PVAF) and word length (β = -.104, t = -2.02, p = .056; see Figure 1.5). However, in our analysis of critical target words, we found a significant positive relationship between the activation of IC1 and character length (main effect of character length, β = .07, t = 2.73, p = .007), and that this relationship depended on whether the word was actually displayed (character length × word presence, β = -.09, t = -2.97, p = .003; See Figure 1.6A). Follow-up models indicated that the positive relationship between IC1 activation and character length held for omitted words (β = .07, t = 2.82, p = .006; Figure 1.5a), but not presented words (p = .37).

Independent Component 2

We did not observe any relationship between the activation of IC2 (4.0 PVAF) and character length, word presence, or their interaction, in analyses of either critical or non-critical words (all p's > .13).
Independent Component 3

In analysis of all non-critical words, we found that the size of IC3 (2.6 PVAF) activation was positively related to character length ($\beta = .013$, $t = 4.66$, $p < .001$; Figure 1.5). In analysis of critical target words, we did not find any effects of Word Length or Word Presence (all $p$’s > .23). However, given the reliable correlation between IC3 amplitude and non-critical Word Length, we conducted analyses of the relationship between IC3 activation and Word Length for omitted and presented words separately. These did not indicate a significant positive correlation between character length and IC3 activation for either presented or omitted words ($p$’s > .12).

Independent Component 4

Because we found that the activation of IC4 (2.0 PVAF) peaked earlier than the other components of interest, we analyzed its relationship to character length at two time windows: 120 msec. to 160 msec. (as for all other analyses) and 100 to 140 msec. (capturing the component’s peak at about 130 msec.). At 120 to 160 msec., we found no relationship between character length and IC4 (all $p$’s greater than .06). However at 100 to 140 msec. we did observe a positive relationship between Character Length of all non-critical context words and IC4 activation ($\beta = .020$, $t = 5.91$, $p < .001$; Figure 1.5). In analysis of critical target words, we did not find any effects of Word Length or Word Presence (all $p$’s > .12). However, given the reliable correlation between IC4 amplitude and non-critical Word Length, we conducted analyses of the relationship between IC4 activation and Word Length for omitted and presented words separately. These analyses indicated a marginally positive relationship between IC4 activation and word length for presented words ($\beta = .027$, $t = 2.29$, $p = .023$; Figure 1.5) but not omitted words ($p = .084$). See figure 1.6B.
Figure 1.5: Relationship between Character Length of all non-target words and activation of components 1, 3, and 4 (left to right). X-axes are word length, y-axes are component activation. The radii of the red circles are proportional to the number of observations at that Character Length. Small black circles are the average component activations for each participant at each Character Length. Bands about the regression lines are 95% confidence intervals. Scalp projections are the same as figure 2. A) There was no significant relationship between the lengths of non-target words and the activation of component 1, further suggesting that it uniquely reflected predictive processes. B & C) Both components 3 & 4 showed a significant positive correlation between non-target word length and activation, confirming the same relationship observed for presented target words.
Figure 1.6: Relationship between the lengths of omitted target words and activation of IC1 (Panel A) and between lengths of presented target words and activation of IC4 (Panel B). The left half of each panel shows event-related activation time series of the IC averaged across short target words (3-6 characters, black curves) and long target words (7-14 characters, red curves). The right half of each panel shows the linear relationship between Character Length and the IC’s average activation in a time window centered on the IC’s first major peak (120-160 msec. post-stimulus onset for IC1 and IC3 and 100-140 msec. for IC4). Each black circle represents one participant’s average activation for words at that length. Each red circle is centered on the mean across participants, with radius proportional to the number of observations at that word length. Bands about the regression lines are 95% confidence intervals. Y-axes of component time series are component activation. Shaded areas indicate the time-windows used for statistical analyses. Inset scalp projections of each component are the same as Figure 2. A) Component 1. Activation for component 1 from 120-160 msec. was more positive for longer omitted words than shorter omitted words. There was no relationship between the activation of component 1 and the lengths of presented words. B) Component 4. Activation for component 4 from 100-140 msec. was more positive for longer presented words than shorter presented words. There was no relationship between the activation of component 4 and the lengths of omitted words.

2.3. Study 1 Discussion
We tested the hypothesis that language comprehenders predict the low-level perceptual features of upcoming words. While we recorded EEG from the scalp, participants read sentences appearing one word at a time on a computer screen, each framed by a vertical crosshair. In half the sentences, a semantically predictable word was omitted from display, leaving only the crosshair. Empty crosshairs (with words omitted) elicited perceptual brain responses between 120 and 160 msec. after the onset of the stimulus that correlated with the character length of the omitted words. This was true when examining either the P1 scalp ERP or a latent component extracted via Independent Component Analysis. Both ERP and ICA effects of omitted words had left occipital-temporal scalp distributions. Because there was no bottom-up linguistic stimulus when a word was omitted from display, the correlation must reflect top-down predictions about the word’s form. Furthermore, because word length and meaning are at best loosely connected, the correlations we observed here are likely to reflect perceptual, rather than semantic representations. Such a conclusion is further supported by the effect’s occipital-temporal scalp distribution, which suggests generators in visual cortex.

The lengths of presented words correlated with two different independent components, which had scalp projections concentrated over right-posterior electrodes. The dissociable pattern of correlations for omitted and presented words suggest separable processes for prediction and bottom-up processing. Such a conclusion is also consistent with the pattern of scalp ERP effects, where P1 amplitude over left hemisphere electrodes correlated with the lengths of omitted words, while the lengths of presented words correlated with the P1 over right hemisphere electrodes. These correlations between word-length and P1 amplitude also mirror the results of previous studies (e.g. Hauk, et al., 2006).

2.3.1 The timecourse of predictions.
The temporal resolution of EEG makes it an indispensable tool for illuminating predictive
dynamics in the visual system. This is because top-down effects might, for instance, occur only in a later
stage of re-analysis, especially in cases where a stimulus is ambiguous or embedded in noise. In the
current study, the latency of the EEG effects indicate that top-down predictions play a role in the earliest
stages of visual analysis.

The omitted word experimental paradigm provides particularly constraining evidence that the
observed brain responses involve prediction, as they cannot be explained in terms of rapid feedforward
semantic access. This resolves a crucial ambiguity about prior reports of early EEG/MEG responses to
unexpected words (Dambacher et al., 2009, Dikker et al., 2010, Kim & Lai, 2012). Consistent with the
current study, these previous results may have reflected a conflict between the sensory input and
predicted perceptual representations for specific words. However, they are also consistent with fast
bottom-up access of lexical-semantic representations (Amsel et al., 2013), leading to conflict between
expected and actual semantic features. Unlike previous research, the present findings cannot be
explained by the latter account, as no stimulus was present to drive bottom-up semantic access.

2.3.2 What would a rational predictor predict?

It has been increasingly common to assume that predictions reflect the operation of a rational
system that is optimally adapted to the statistics of its environment. This leads to critical questions
about the types of predictions that such a system would deem worth computing. It could be the case,
for example, that some types of predictions are not made because they are likely to be wrong, and
computing them would therefore result in a metabolic net-loss (Kuperberg & Jaeger, 2016). This might
apply to the pre-activation of specific word forms, as language’s inherent variability (Levy, 2008; Jaeger
& Snider, 2013) could make it difficult to make consistently accurate predictions. In contrast, predictions
of high-level semantic features may be more accurate on average, because they can be made without
commitment to one particular word. For instance, when processing “For several years, the McCanns had lived in the same …”, prediction of a semantic feature like LIVABLE-PLACE might be highly accurate, consistent with multiple continuations (HOME, HOUSE, NEIGHBORHOOD, TOWN). Our results provide an important empirical constraint on these issues, as they imply that low-level prediction of word-forms can in fact be a rational strategy in some contexts.

Furthermore, our results suggest that low-level form prediction may represent the stronger notion of pre-activation, which can be distinguished from more passive forms of prediction (Van Petten & Luka, 2012, Kuperberg & Jaeger, 2016). To show how prediction does not necessarily imply pre-activation, consider the phenomenon of priming. Given a prime word (“doctor”), it is typical that both reaction times and brain responses (e.g., the N400) are reduced when the subsequent target is somehow related to the prime (“nurse”) than when it is unrelated (“avocado”). Priming effects are often explained by appealing to the “spreading of activation” (Collins & Loftus, 1975) between related concepts and words, so that the recognition of “doctor” automatically facilitates the retrieval of features related to “nurse.” Critically, however, this spreading of activation is not equivalent to a pre-activation of the complete meaning of the target word – only the features that incidentally overlap between the prime and target are pre-activated, and these will rarely, if ever, be completely overlapping. One feature of the target that would not be pre-activated by spreading activation would be its orthographic form (N-U-R-S-E), as this constitutes a largely arbitrary feature of a word or concept. Because our data support the predictions of word lengths (as an arbitrary aspect of orthographic representation), they support the stronger sense of word form prediction as pre-activation, rather than incidental spreading of activation.

Finally, it may be that low-level pre-activation of this sort is only licensed when predictive affordance is strong enough. Our particular paradigm, for example, involved the use of sentences that provided strong semantic support for one particular word (cloze values > 85%), and a task where
participants were instructed to anticipate upcoming language. Many real world situations possess such characteristics, either because language comprehenders frequently deal with topics with which they are familiar, or because noise in the environment makes prediction a useful strategy. However, some contexts are rich with novel information and surprising language, and thus SH4 and study 3 are meant to address the issue of whether such low-level predictions occur in less constraining linguistic contexts.

2.3.3 Prediction or prediction error?

Our results are compatible with theories proposing that the hierarchical organization of the brain reflects its fundamental reliance on predictions and prediction error (e.g. Clark, 2013; Rao & Ballard, 1999). Predictive coding models argue that top-down connections between cortical areas carry predictions about future inputs, while bottom-up connections carry information about the discrepancy between what was predicted and what was observed. The feed-forward error signal is then used to adjust connections throughout the system, tuning it in preparation for future stimuli.

One potential link between our effects and such models is that the correlation between the length of omitted words and our left hemisphere EEG signal may reflect the generation of top-down predictions. This might occur via preparation of retinotopic V1 cells that have receptive fields overlapping with the word’s expected horizontal extent. Alternatively, the effect could also represent the prediction error between the expected word length and the actual physical stimulus (i.e. the vertical crosshair). Both of these explanations are consistent with the model of Rao & Ballard (1999), but are equally plausible given our effect pattern. Thus, it would be premature to propose a one-to-one mapping of our results to the specific components of a predictive coding model. Further studies could resolve this issue by orthogonally manipulating the discrepancy (in number of characters) between semantically predictable words and the length of replacement non-word strings, which would allow one to tease apart aspects of the EEG signal that are related to predictions and prediction error, respectively.
To clarify whether the effects we observed in this study reflect prediction generation or prediction error, we designed Study 2 as an explicit manipulation of prediction error, as described below.

3. Study 2: Dissociating Prediction from Prediction Error

The results from Study 1 strongly suggest that language comprehenders can make low-level predictions about upcoming word forms, as brain responses to omitted words still correlated with the expected word’s length. As noted, however, the results do not allow us to discriminate between whether the effect represents top-down prediction of the expected length, or the prediction error between the expected and perceived stimulus. This is because the magnitude of the expected character length and the difference between the expected character length and the empty crosshair stimulus are essentially equivalent.

As a result, a follow-up study was designed that would allow us to differentiate between these prediction-generation and prediction-error hypotheses. Rather than replace the critical word with the empty crosshairs stimulus, we instead replaced the critical word with a string of symbols that varied in length. In particular, we fixed the lengths of the replacement stimuli to have a near-0 correlation with the lengths of the expected words, so that the length of the stimulus that was actually seen could not be predicted on the basis of the expected word’s length.

This manipulation allows us to again examine effects of the expected word length (i.e. effects of prediction) as well as the effects due to the discrepancy between the perceived and expected character lengths (i.e. prediction error) (SH2). Both factors can then be simultaneously used to predict the amplitude of the resulting visual components, in order to disentangle which aspects of the EEG may be related prediction and prediction error, respectively. Furthermore, we can exploit the temporal resolution of EEG to examine whether prediction precedes the computation of prediction error, which
would provide further evidence that predictive coding mechanism contributes to visual word recognition (SH3).

In summary, study 2 had two principle goals. The first was to establish a clearer relationship between early visual ERP components and the error signal that might occur when a word’s form deviates from predictions, as this question could not be answered directly by Study 1. Second, because the design of Study 2 allows us to simultaneously model the effects of both predicted word length and prediction error, we sought to establish that predictions are reflected in the EEG signal prior to prediction error, as some predictive coding models imply should be the case (e.g. Rao, 1999).

3.1 Study 2 Methods

3.1.1 Participants and Stimuli.

30 University of Colorado undergraduates participated in Study 2. All participants were right handed and had normal or corrected-to-normal vision. One participant’s data was excluded due to less than 70% of trials being retained after artifact rejections and incorrect behavioral responses. This left 29 subjects for data analysis. All participants were compensated with either course credit or 20 USD.

The same set of 200 sentences from Study 1, plus an additional 50, were used as stimuli for Study 2. The 50 additional sentences were taken from the cloze norming study completed prior to Study 1, and included stimuli whose target word cloze probabilities ranged from 70-85%. Overall, the average cloze probability for the sentences in Study 2 was 91% (SD: 9%). The mean target word length was 6.67 characters (range 3-14, SD: 2.54).

Eight presentation lists were constructed containing one version of each stimulus sentence. For each presentation list, 80 sentences occurred with the expected word displayed unaltered. The other 170 stimulus sentences had their critical word replaced by a non-alphanumeric string of between 1 and
16 characters long. Replacement strings were generated by randomly sampling from the list of symbols below, with replacement.

Within each presentation list, the lengths of replacement stimuli were chosen such that their correlation with the lengths of the to-be-replaced target words would be as close to 0 as possible, to prevent participants from being able to accurately predict the length of the symbol string that might replace an anticipated word. Across all lists, this correlation was .004. This was done while simultaneously respecting the following constraints: (1) the lengths of the replacement strings chosen for each critical word were different in each list, (2) the difference between lengths of the shortest and longest replacement strings for each critical word was at least 12, and (3) the lengths of the replacement strings chosen for each critical word differed in length from each other replacement string for that word by at least 2 characters. Criteria 1-3 were chosen to ensure a similar distribution of replacement string lengths for each target word length, while simultaneously respecting the constraint that the correlation between target word lengths and replacement string lengths was close to 0. For example, 4 character long target words had approximately the same distribution of replacement string lengths as 8 character target words.

3.1.2 Procedure.

EEG recording took place using the same apparatus as Study 1. Sentence stimuli were presented one word at a time in Courier New in the center of a black screen, in this case without the crosshair used in Study 1. Each character occupied ~.8 degrees of visual angle. We used the same stimulus timing as Study 1: 300 msec. followed by a 300 msec. blank screen (SOA = 600 msec.). In 170 of the 250 stimulus sentences, a predictable word was replaced by a non-word character string (see stimulus section, above, and Figure 2.1). After each sentence, a screen appeared asking participants whether they noticed if a
word was replaced. If they answered yes, they were asked to name what word they thought was replaced by speaking into a microphone. Answers were coded online by experimenters, who listened to participants through speakers while sitting outside of the experiment booth. Participants saw 6 example stimuli (4 sentences with the critical word replaced and 2 sentences with the critical word presented) prior to starting the experiment. An experimenter remained in the room with the participant to answer any questions during the practice period. Sentence stimuli were presented in five blocks of 50 sentences each, with a break between blocks.

Figure 2.1: The task for study 2 was similar to study 1, except that (1) the target word was now replaced by a variable-length non-word character string in 70% of cases, and (2) the task for the participants was to identify if a word had been replaced, and if so, what word they thought was replaced.
3.1.3 Data Analysis

Defining Prediction Error

We manipulated the discrepancy, in number of characters, between the lengths of predictable sentence-embedded target words and stings of replacement symbols. Given this approach, there are several ways in which we could have operationalized the measure of prediction error to be used in our statistical analyses. We ultimately decided to quantify prediction error by taking the absolute value of the difference between the lengths of the replaced words and replacement strings as our measure of prediction error. This means that large deviations from the replaced word’s length will lead to larger values of prediction error, regardless of whether the error represents an over- or under-estimation of the expected word’s length. We chose this method primarily because it helps disentangle effects of prediction error size from effects due to the length of the predicted (but replaced) word, as the correlation between the two measures would otherwise be perfect if we had maintained the sign of the prediction error (e.g., treated errors of +2 and -2 characters as different). Using the simple (i.e., signed) difference between the lengths of the replacement strings and the lengths of the replaced target words would stymie our efforts to statistically isolate the effects of prediction error from predicted word length through multiple regression. To see this, consider a regression analysis containing two predictors of the EEG signal, one being the length of the replacement string (i.e. the length of the stimulus actually seen by the participants), and the other being the simple-difference error measure. Any effect of the simple difference measure would then be perfectly confounded with the effect of the length of the replaced word, since the measure is derived from the difference between the replacement string and replaced target word. Put another way, let us define $R$ as the length of the Replacement string, $T$ as the length of the replaced Target word, and $E$ as prediction Error, itself defined as $E = T - R$. First, because $E$ is derived from $T$ and $R$, all three variables cannot be included as simultaneous predictors in a multiple
regression (the model would be rank deficient). If we then decided to include just \( E \) and \( R \), any effect of \( E \), controlling for \( R \), would be perfectly confounded with the effect of \( T \). Defining \( E = |T - R| \) removes the perfect confounding, as for any \( E \), there are now two (rather than 1) potential values of \( T \). This allows all three measures to be used as simultaneous predictors in a multiple regression.

Using the absolute value of the difference between target word lengths and replacement string lengths as our measure of prediction error involves making the assumption that, when an error is detected, the magnitude of the error is computed. Importantly, there may be a stage of processing during which the sign of the error (i.e., whether a prediction was an over- or under-estimate) does matter – it may be required, for instance, to know the direction in which future predictions may need to be adjusted. However it is equally true that error magnitude is will also have to be represented in some way, and that equally large under- or over-estimations will require roughly the same amount of energy to represent. However, because the goal of study 2 is to dissociate prediction generation from prediction error, it will be necessary for us to rely on error magnitude as our measure. Doing so allows us to simultaneously model the effects of both prediction and prediction error on the visual EEG response, while each measure serves as a covariate for the other.

**Statistical Analyses**

Only trials in which a word was replaced by a non-English string were included in the analyses reported below (control trials where the expected word was actually displayed were excluded).

Study 2 was designed in part to determine whether the generation of word-form predictions precedes the computation of prediction error. This required testing for effects of both variables at consecutive time-intervals with a fine enough temporal resolution that any such differences in timing might actually be observed. Towards this end we split the interval of 75-225 msec. post-stimulus into 6
smaller intervals of 25 msec. (or 5 samples of EEG) each, and computed the same type of mixed-effects regression analysis on each interval separately.

Regression analyses included the following three fixed effects: (1) length of the replaced word (used as a measure of word-form prediction), (2) length of the replacement string that was actually displayed (used as a measure of bottom-up visual processing) and (3) the absolute difference between the first two measures (used as our measure of prediction error), hereafter referred to simply as prediction error. All models included random effects for participant (random slopes and intercepts) and stimulus item (random intercepts). The dependent measure in all cases was the mean voltage in each time interval.

Because this approach involved the repeated testing of similar statistical models, we elected to analyze a single group of electrodes broadly covering the occipital part of the scalp, to minimize further multiple comparisons issues that would arise from tests of multiple scalp locations. This electrode group included channels OZ, O1, O2, CB1, CB2, PO7 and PO8.

Finally, although we consider these regression analyses to be planned tests, we acknowledge that executing the same analysis for six consecutive time intervals may lead to an increased chance of false positives. We thus submitted our results to the Holm-Bonferroni method (Holm, 1979) to control the family-wise error rate. The method was applied separately for each predictor variable, using an overall alpha level of .05. The critical value for each test, and whether or not \( H_0 \) was rejected for that test, is shown in table 3.1.

3.2 Study 2 Results

Behavioral Results
Of sentences in which the critical target word was replaced by a symbol string, participants on average named the intended word 87.5% of the time (range .66 - .97; SD = .07). Only trials where participants correctly named the replaced word were included in the analyses reported below.

Regression Results

Below we summarize the results of the regression models described above, starting at the first time interval (75-100 msec. post-stimulus). See table 3.1 for the full complement of results. The Holm-Bonferroni critical values for the tests described below are denoted in brackets.

From 75-100 msec. post stimulus onset we found no relationship between changes in voltage and either replacement string length, replaced word length, or prediction error (all p’s > .51).

From 100-125 msec. (including the rising slope of the P1 ERP), we found a marginally negative relationship between EEG voltage and the length of the replaced word ($\beta = -.10, t = -1.91, p = .058 [.013]$), and a marginally positive relationship between voltage and prediction error ($\beta = -.088, t = -2.10, p = .036 [.036]$). See figure 2.3.

From 125-150 msec. (including the peak of the P1 ERP), we continued to find a marginally significant negative relationship between voltage and the length of the replaced word ($\beta = -.013, t = -2.63, p = .009 [.008]$). We also found voltage to become more negative as the length of the replacement string increased ($\beta = -.060, t = -2.21, p = .027 [.017]$). See figures 2.2 and 2.3.

From 150-175 msec. (including the beginning of the rising slope of the N170), we found that voltage became marginally more negative as the length of the replaced word increased ($\beta = -.093, t = -1.91, p = .058 [.010]$). The negative relationship between replacement string length and voltage became much stronger at this time interval ($\beta = -.30, t = -11.35, p < .001 [.008]$). We also found that as prediction
error increased, voltage became marginally more positive (β = .090, t = 2.32, p = .021 [.010]). See figures 2.2 and 2.3.

From 175-200 msec., the negative relationship between replacement string length and voltage remained strong (β = -.32, t = -12.25, p < .001 [.010]). We also found a positive relationship between prediction error and EEG voltage at this time interval (β = .17, t = 4.23, p < .001 [.008]). See figures 2.2 and 2.3.

From 200-225 msec. (including the peak of the N170), we continued to observe the negative relationship between replacement string length and voltage indicating larger N170 amplitude (β = -.21, t = -7.94, p < .001 [.013]), as well as a marginally significant positive effect of prediction error (β = .068, t = 1.71, p = .087 [.017]). See figure 2.2.
<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Predictor</th>
<th>β</th>
<th>t-value</th>
<th>p</th>
<th>Holm-B CV</th>
<th>Sig. under Holm-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>75-100 msec.</td>
<td>Replacement String Length</td>
<td>-0.007</td>
<td>-0.27</td>
<td>0.787</td>
<td>0.050</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.016</td>
<td>-0.34</td>
<td>0.735</td>
<td>0.025</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>-0.026</td>
<td>-0.65</td>
<td>0.517</td>
<td>0.050</td>
<td>N</td>
</tr>
<tr>
<td>100-125 msec.</td>
<td>Replacement String Length</td>
<td>0.028</td>
<td>1.00</td>
<td>0.318</td>
<td>0.025</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.097</td>
<td>-1.91</td>
<td>0.058</td>
<td>0.013</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>-0.088</td>
<td>-2.10</td>
<td>0.036</td>
<td>0.013</td>
<td>N</td>
</tr>
<tr>
<td>125-150 msec.</td>
<td>Replacement String Length</td>
<td>-0.060</td>
<td>-2.21</td>
<td>0.027</td>
<td>0.017</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.125</td>
<td>-2.63</td>
<td>0.009</td>
<td>0.008</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>-0.046</td>
<td>-1.16</td>
<td>0.245</td>
<td>0.025</td>
<td>N</td>
</tr>
<tr>
<td>150-175 msec.</td>
<td>Replacement String Length</td>
<td>-0.298</td>
<td>-11.35</td>
<td>0.000</td>
<td>0.008</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.093</td>
<td>-1.91</td>
<td>0.057</td>
<td>0.010</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>0.090</td>
<td>2.32</td>
<td>0.021</td>
<td>0.010</td>
<td>N</td>
</tr>
<tr>
<td>175-200 msec.</td>
<td>Replacement String Length</td>
<td>-0.323</td>
<td>-12.25</td>
<td>0.000</td>
<td>0.010</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.067</td>
<td>-1.48</td>
<td>0.141</td>
<td>0.017</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>0.166</td>
<td>4.26</td>
<td>0.000</td>
<td>0.008</td>
<td>Y</td>
</tr>
<tr>
<td>200-225 msec.</td>
<td>Replacement String Length</td>
<td>-0.212</td>
<td>-7.94</td>
<td>0.000</td>
<td>0.013</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Replaced Word Length</td>
<td>-0.008</td>
<td>-0.17</td>
<td>0.862</td>
<td>0.050</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Prediction Error</td>
<td>0.068</td>
<td>1.71</td>
<td>0.087</td>
<td>0.017</td>
<td>N</td>
</tr>
</tbody>
</table>

**Note.**

Holm-B CV: critical value for the test under the Holm-Bonferroni Method

Table 3.1: Results of regression analyses for study 2. The far right column indicates whether the results is significant under the Holm-Bonferroni method.
Figure 2.2 Left: ERP effect of character length of replacement string at occipital electrodes (unless otherwise noted, all subsequent ERP traces will be from the same scalp electrodes OZ/1/2, POZ/7/8, CB1/2). For display purposes only, data are split into quartiles (all predictor variables were treated as continuous in the regression analyses that are reported above). The grey box indicates the time-window in which the linear effect of replacement string length is significant at $p = .05$. Top right: scalp distribution of t-statistics associated with the character length parameter estimate (slope), when replaced word length and error size are controlled for. Top of scalp plots is the front of the head, bottom is the back of the head, left is left and right is right. Blue values indicate negative slopes, red values indicate positive slopes (this is true for all further scalp plots). Each scalp plot is the average over 25 msec. of time, starting at 75 msec. (top left) and ending at 225 msec. (bottom right). The interval above each scalp plot indicates the time interval. Bottom right: scatter plot of relationship between character length and voltage, 150-225 msec., when controlling for error size and replaced word length. On the y-axis are the residuals from a model including replaced word length and error size, on the x-axis is word length.
Figure 2.3 Top Half: ERP effect of length of replaced word. Data are split into quartiles for display purposes only. The data show that longer replaced words led to smaller P1 amplitudes. Top right: scalp distribution of t-statistics associated with the replaced word length parameter estimate (slope), when replacement string length and error size are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 100 msec. (top left) and ending at 200 msec. (bottom right). Inset: scatter plot of relationship between character length of replaced word and voltage, 100-150 msec., when controlling for error size and replacement string length. On the y-axis are the residuals from a model including replacement string length and error size, on the x-axis is replacement word length.

Bottom Half: ERP effect of length of prediction error (absolute value of difference between replacement string length and replaced word length). The data show that larger prediction errors led to more positive. Top right: scalp distribution of t-statistics associated with the error size parameter estimate (slope), when replacement string length and replaced word length are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 100 msec. (top left) and ending at 200 msec. (bottom right). Inset: scatter plot of relationship between error size and voltage, 150-200 msec., when controlling for replaced word length and replacement string length. On the y-axis are the residuals from a model including replacement string length and replaced word length, on the x-axis is error size.

3.3 Study 2 Discussion and Summary
Study 2 addressed two related issues: (1) whether the P1 effect observed in Study 1 primarily reflected prediction error or prediction generation, and (2) whether a serial ordering of word-form prediction and prediction error can be established in the visual EEG signal. We briefly discuss each of these issues in turn, returning to them in more detail in the General Discussion.

3.3.1 Effect of Prediction Error on the Visual EEG Response

We designed study two primarily as a manipulation of visual prediction error, by replacing highly predictable words within a sentence with a non-word symbol string that was either longer or shorter than the replaced word. When prediction error was quantified as the absolute difference between the lengths of the replaced word and replacement string, we found a brief window of activity (150-200 msec. post-stimulus) in which there was a positive correlation between prediction error size and EEG voltage, when controlling for both the lengths of the replaced word and the length of the replacement string (see figure 2.3, bottom). Because this relationship was positive, and partially overlapping with the window used to analyze the P1 effect in Study 1, our results support the possibility that the prior results represent the computation of prediction error, rather than prediction generation. The case becomes marginally stronger when one considers that the association between the replaced word’s length (which is an analogous measure to the length of the omitted word in Study 1) and the EEG during the peak of the P1 (100-150 msec.) was negative (figure 2.3, top).

3.3.2 Timing of Prediction, Bottom-Up Processing, and Prediction Error

Study 2 had two primary goals. First, we sought to understand whether the primary finding of Study 1 (increasing P1 amplitude in response to longer omitted words) may have been due to prediction generation itself, or rather the prediction error between the length of the predicted word and the actual stimulus. The results of Study 2 are most compatible with the latter possibility, as we found a positive effect of prediction error on the later part of the P1 component, while the length of the predicted (but
replaced) word had an earlier negative effect on P1 amplitude (the opposite of what we found in Study 1).

The second goal of Study 2 was to establish whether the brain response to a word reflects word-form prediction before it reflects prediction error. Establishing this sequence of events is critical to determining whether the architecture for word recognition is truly predictive, as top-down predictions must logically precede the computation of prediction error. We found some support for this sequence of events, as we observed that from 125-150 msec., the signal appeared to switch from representing the length of the expected word to representing the length of the (unexpected) replacement string. Then, from 150-200 msec. we observed simultaneous – but opposing – effects of replacement string length and prediction error. Nevertheless, we should note that the negative effect of replaced word length just barely missed the cut-off for statistical significance under Holm-Bonferroni (p = .009 when alpha = .008).

This timeline of effects provides some initial evidence that the brain’s response to a visual word is divided into distinct, but rapid stages of processing. Initial stages may be dominated by the system’s “best guess” of what the visual form of the word at the current moment should be, followed by a comparison of this prediction to the information actually arriving through sensory channels. This comparison appears to result in a brief representation of prediction error, with larger errors leading to more positive voltage.

Overall, the results of Study 2 mark an important step towards establishing whether top-down predictions, and any resulting errors, are computed about even the low-level (i.e., size) characteristics of upcoming words. However, the extent to which these results are generalizable to most linguistic contexts can still be questioned, as both studies 1 and 2 examined brain responses to words that were highly predictable (at least 70% cloze probability). This leaves open the possibility that the prediction-related effects we have seen thus far do not form a core part of the word recognition machinery, and
instead reflect a mere “strategy” deployed in cases where prediction is likely to be successful. If this were the case we would have to reject the possibility of predictive coding being a fundamental part of word-form recognition, as top-down predictions are an automatic, core component of such models. Study 3 is therefore designed to test for effects of prediction and prediction error at every word read by the participants in studies 1 and 2. If we can find evidence for form prediction even in cases of low predictive affordance, then the case for a fundamentally predictive model of word-form recognition grows considerably stronger.

4. Study 3: Testing for consistent form-level prediction

Studies 1 and 2 represent an experimental approach to investigating the extent of word-form prediction and prediction error computation during reading. Like all experiments, they sacrifice some generalizability for the sake of careful control over the stimuli. Of particular concern is the fact that both studies were only designed to test for prediction effects in highly predictable (i.e. high cloze) contexts, leaving open the possibility that word-form prediction only occurs in such cases. One could argue, for instance, that language’s inherent unpredictability could make the average cost in prediction error outweigh the benefits that are accrued in the (relatively infrequent) cases of high predictability. In such a case it may be that predictions are made opportunistically, rather than a matter of course. In turn, this would strongly suggest that form-level prediction is closer to a “feature” of the word recognition system, rather than a fundamental process.

On the other hand, it can also be convincingly argued that consistent prediction might be a net-plus metabolically, as language’s grammatical regularities ensure that there is never a sentence context in which the probability distribution over the next word is uniform. The case become even stronger if
one is willing to entertain that prediction may be a weighted expectation of many different outcomes, rather than a firm commitment to any particular one.

But because we cannot precisely estimate the computational or metabolic costs of prediction error, we cannot give an \textit{a priori} answer to whether the language system is \textit{constantly} or only \textit{opportunistically} predictive. We must therefore look for empirical evidence of prediction in language contexts that do not strongly support the prediction of one particular word (as was strictly the case in studies 1 and 2) (SH4). To do this properly we must have an estimate of the probability distribution, at each sentence position, over what the form of the next word will be. To obtain this estimate for the purposes of study 3 we took the approach of training a neural network model to provide a probability distribution over the identity of each stimulus word, given the words that occurred prior to it in the sentence. By training the model on a large enough corpus of text, it can be used as a proxy for the distributional knowledge that participants would likely incorporate into their own word-by-word predictions. Critical to our purposes, such information also allows us to derive a measure of form-specific prediction, using the lengths of the words that the model most strongly predicts should occur at each sentence position. Furthermore, a measure of word-form prediction \textit{error} can also be derived, by comparing the length of the word that was actually displayed at each sentence position, to the model-derived predicted word length. We outline the details of this approach below.

4.1 Methods

Our goal is to develop an estimate what our human subjects may have been predicting at each non-target position of Study 1 and Study 2’s sentence stimuli. Because we did not cloze-norm these sentences positions (and doing so would require an impractical amount of data collection), it is necessary to develop a model that can generate predictions about each upcoming word \(w\), given the previous \(w-1\ldots w-n\) words. More specifically, for each word input to the model, a probability distribution
over the identity of the next word w+1 should be output. This probability distribution can then be used to generate predictions about the length of w+1, based on lengths of the words that the model predicts are most likely to continue the sentence.

While there are many ways to define a predictive language model, the current approach combines two pre-existing machine learning methods for word representation and sequence learning, respectively. To represent each word at the model’s input we use GloVE, a learned distributed representation in which each word in a vocabulary is encoded as a vector of length 300. To represent the statistical dependencies between words, and thus be able to produce predictions, a recurrent neural network using the long short term memory (LSTM) architecture will be employed. Below I provide a brief summary of the GloVE and LSTM models in general, and then proceed to describe the particular architecture and training procedure of the model used in the current study.

4.1.1 LSTM model for word prediction

LSTM is a variety of recurrent neural network (RNN) that was explicitly designed to learn long distance dependencies between elements of sequence (Hochreiter & Shmidhuber, 1997). This is a particularly important part of language comprehension, as two co-dependent words in a sentence may be separated by many other words (consider “The man who needs to get a job after graduating from school last May plans to make some calls”; here the verb “plans” depends on the singular form of “man” (if “man” was “men” the verb would have to be “need”).

The performance of LSTM models in learning sequences often surpass those of traditional recurrent networks that simply combine the new input at each time step with the hidden state of the network at the previous time step. In such models the influence of prior hidden states on the processing of the current input falls off as the separation (in time) between the current input and past input increases. The LSTM architecture avoids this problem by implementing a system of three “gates” that
control what prior information influences processing of the current sequence element (e.g., word), and what information is passed on to the next state. Like classic RNNs, the previous hidden state of the network is allowed to influence the current state, but is also subject to the operation of the gating system which (1) learns what information from the previous state to “forget” (2) learns what information from the current input (e.g. current word) should be considered when computing the output and (3) what parts of the current state should be passed on to the next state. Thus, if given enough training samples, the network can learn to “hold on” to information for as long is needed to generate the correct output.

In many language applications, including the current one, LSTMs are trained to predict the next word in a sentence, given the previous words. Thus, the input to the network is a representation of the current word \(w\) (in this case the GloVE vector), and the output is some estimate of the next word \(w+1\), usually (and as in our case) in the form of a probability distribution over some vocabulary of a fixed size.

4.1.2 GloVE Vectors for Word Representation

To represent words as they are input to the LSTM, we use a pre-trained version of the GloVE model (Global vectors for word representation; Pennington, Socher & Manning, 2014). After training, GloVe represents each word in the training corpus as an arbitrarily long vector (50 in the current case). Training happens on a global word-word co-occurrence matrix populated by moving a window of arbitrary size across a large corpus (a common choice for window size is 15 words long). The resulting vectors are trained with the objective that the dot-product between any two of them is close to the log of the co-occurrence probability of their corresponding words \(i\) and \(j\). This co-occurrence probability is in turn equal to \(X_{ij}/X_i\) or their joint entry in the co-occurrence matrix \(X\), normalized by the total number of times that any word occurs in the context of \(i\). Because the logarithm of \(X_{ij}/X_i\) is equal to \(\log(X_{ij}) - \log(X_i)\), the information reflected in the co-occurrence probabilities gets encoded as differences in the
vector space. The resulting vectors end up representing some subtle properties of word meaning, owing to the fact that ratios of co-occurrence probabilities can be surprisingly informative, as illustrated by the following example given by Pennington, Socher & Manning (2014). Suppose we want to capture the difference in meaning between the word “ice” and “steam”. One way to do this is compare their relative co-occurrence with certain other words in the language. For example, the probability of “ice” given “solid” is 1.9 x 10^{-4}, while the probability of “steam” given “solid” is 2.2 x 10^{-5}. The ratio of the former probability to the latter is 8.9, whereas the same ratio using a probe word unrelated to either “ice” or “steam” will be close to 1 (since “ice” and “steam” are equally likely to co-occur with, for example, “fashion”). After training, this information gets translated into the GloVe vector space in terms of differences – for example, the cosine similarity between two vectors which encode the differences between (steam – ice) and (gas – solid) is very near to 1 (indicating a high degree of semantic similarity).

Because each GloVe vector captures potentially subtle information about the word’s meaning, and reflects similarities between words that are semantically related, they are superior to a localist encoding method for the LSTM input. Whereas a localist encoding (unique, orthogonal vectors for each word) of the input word to the LSTM would only allow the network to learn co-occurrence information between words, using the GloVe vector representation will allow the model to make subtler predictions about upcoming words (for example, to predict with some confidence “libation” given “pour me a …”, even if the model never encountered a co-occurrence of “pour’ and ‘libation”).

4.1.3 Model Architecture

A high-level representation of the neural network model used to generate word-by-word predictions for our sentence stimuli can be seen in figure 3.1. The model is fed words one at a time and generates a probability distribution over the next word as its output.
Figure 3.1 Schematic of neural network model used in study 3.

The model includes two input layers. The first encodes the part of speech (POS) of the input word as a binary vector of length 46, one unit for each of the possible POS tags in the penn treebank tagset. POS tags were obtained by tagging the training corpus and sentence stimuli with the Stanford tagger. The POS input uses a ‘one-hot’ encoding scheme, where the vector element corresponding to the correct POS tag is set to 1 and all other elements are set to 0. The other input layer represents the lexical identity of the input word as its length-300 GloVE vector obtained from a vocabulary of pre-trained vectors provided by Pennington, Socher & Manning (2014). During training, the weight matrix representing the GloVE vectors is not frozen, so that the word embeddings can be tweaked to better represent the idiosyncrasies of the training corpus (the present training corpus differs from the one used
by the original GloVE developers; see below for details). The word embedding layer and POS layers are then concatenated together to provide the input to the first LSTM layer. Because not all words in the training corpus could be included in the model’s vocabulary, as special token was assigned to all “out of vocabulary” (OOV) words. This OOV token was included as a single entry in the model vocabulary and was subjected to the same training procedures as all other words.

Previous work in computational linguistics has shown the benefit of stacking many LSTM layers in a sequential fashion (Graves, Jaitly & Mohamed, 2013). We follow this model here, with the first two LSTM layers containing 1024 units and a third containing 512 units. The number of units and layers mainly reflect consideration of training time and GPU memory limits. In each LSTM layer we employ a dropout rate of .2. Dropout (Srivastava et al., 2014) is a method used to prevent model overfitting, and works by randomly deactivating a subset of hidden units (in this case, 20% of them) during the training of each sequence.

The model output is a typical fully connected neural network layer with 19977 units (one unit for each word in the vocabulary). The softmax function \(\frac{e^{x_i}}{\sum e^{x_i}}\) where \(x_i\) is the activation of the \(i\)th output neuron) is applied to the activation of the output units to squash their values to between 0 and 1, thus yielding a probability distribution over the model’s vocabulary in response to each input word.

The model vocabulary was determined by including all the content words from studies 1 and 2, as well as additional content words from two other unrelated studies from our laboratory. The remaining words were chosen as the most common words (including closed class words and punctuation) in the google 1B word benchmark dataset (https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark), exclusive of the content words already included from the
sentence stimuli. We also included a special beginning-of-sentence token that was prepended to each training sentence, so that estimates about sentence-initial words could be made.

4.1.4 Model Training and Evaluation

The LSTM model was trained on a portion of the Google 1 Billion word benchmark dataset, as well as the complete text of the TASA corpus (Landauer, Foltz & Laham, 1998). The Google portion of the corpus was selected by extracting all sentences in the corpus that contained at least one of the open class words from the stimuli used in studies 1 and 2. This yielded a training corpus of approximately 200M words.

Training was done in batches of 30 sentences (i.e. model weights were updated after every 30 sentence sequences) in epochs of 40,000 sentences with the objective of minimizing the cross-entropy (AKA multi-class log-loss) between the input and output patterns (de Boer, Kroese, Mannor & Rubinstein, 2005). Optimization was done via the ADAM algorithm (Kingma & Ba, 2014). LSTM model development and training was done on the GPU nodes of the University of Colorado Summit computing cluster using the keras (https://keras.io/) and theano (http://deeplearning.net/software/theano/) frameworks for the Python programming language. The model was trained for 105 epochs (approximately 4.2M sentences sampled without replacement from the corpus described above).

After training the model was able to reach a per-word perplexity of 85 when tested on the sentence stimuli from studies 1 and 2 (which were not part of the training set). The model was also qualitatively evaluated by examining the extent to which it generated reasonable predictions about the sentence stimuli in studies 1 and 2. A visual representation of how well a model-derived measure of word length prediction (described in detail below) tracks the actual word lengths of a sample sentence is shown in figure 3.2. Across all sentences, the cross correlation between this model-based measure of length prediction and the actual lengths of the words in the sentence stimuli was .42 (Confidence
Interval = [0.38, 0.46]). This indicates that the model generates reasonable “guesses” about upcoming word lengths but is also “wrong” by enough of a margin in enough cases to generate reasonable prediction errors.

Figure 3.2 Top: Illustration of how the model-derived predictions of word length line up with the actual lengths of the words, for an example stimulus sentence. Red line is the model’s estimate of what the word length at that position should be, quantified as the sum of the length of each word in the model vocabulary multiplied by its probability at that sentence position. The blue line is the actual length of the word at that sentence position. Green line is model entropy at that sentence position. X axis shows the actual stimulus word at each sentence position. Bottom: the model’s predictions for the 20 most likely words at each sentence position, ranked by their probability. The 20 most likely words typically covered over 95% of the probability mass at each position.

OOV = out of vocabulary

4.2 Using the Language Model to Predict Visual EEG Responses
After training, the language model was used to generate probability distributions over the next word \( w+1 \) at each word position \( w \) (except target word positions) in each of the sentence stimuli used in experiments 1 and 2. This probability distribution formed the basis of the measures that were used to predict the visual EEG response (P1/N170 ERP components) to each of the non-target words read by the participants in studies 1 and 2. These measures are discussed briefly below.

The first measure we considered was a sum of the lengths of the words in the model’s vocabulary, with each length weighted by the word’s probability that it occur at each non-target sentence position (\( \sum_i p(v_i)\text{length}(v_i) \) where \( v_i \) is the \( i \)th entry in the model vocabulary). This measure makes the assumption that word-form predictions can be parallel in nature, in the sense that the brain considers multiple possible upcoming forms, weighted by their likelihood. To give a toy example, say that at a particular sentence position, the model strongly predicts specific words that have lengths 6,4, and 2 with probabilities .33,.25 and .2, respectively (with the rest of the vocabulary making up the rest of the probability mass). The predicted word length would therefore be close to \( \sum(.33*3+.25*4+2*.2) = 2.4 \) (plus a small amount extra due the more weakly predicted words).

We also constructed two measures designed to reflect prediction error between the model-based word length predictions described above and the length of the actual stimulus that was displayed to the subject. Error Measure 1 was designed to be comparable to the error measure used in study 2, and was calculated (for each sentence position) as the absolute difference between the value of the model’s predicted word length (described above) and the length of the actual word that was displayed to the participant. Error Measure 2 was the sum of the probability-weighted absolute differences between the length of the stimulus and every word \( v_i \) in the model vocabulary:

\[ \text{Error Measure 1} \]

\[ \text{Error Measure 2} \]

---

3 The complete sentence stimuli, including the intended target words, were used to generate the probability distributions, but only resulting estimates (and corresponding EEG data) for non-target words were actually used in our analyses.
\[\sum_i (|\text{length(stimulus)} - \text{length}(v_i)|) p(v_i).\] Error Measure 2 thus takes into account the difference between the displayed word length and each of the word lengths in the vocabulary, rather than the difference between displayed word lengths and the “overall” expected word length that Error Measure 1 reflects. For both measures, larger values reflect larger deviations from the model’s predictions.

Finally, in order to test whether our measures of word-form prediction and prediction error were related to the EEG signal across all sentence positions - not just those which are highly predictable - we included as additional predictors the interaction between the model’s entropy at each sentence position and the two measures described above. We defined entropy following Shannon (1948) by using the formula \[-\sum_i P(x_i) \log P(x_i)\] where \(P(x_i)\) is the model-estimated probability of the \(i\)th word in model’s vocabulary. We used entropy as a measure of the “predictability” of each sentence position, with high values of entropy meaning the model was very uncertain about the identity of the word at that sentence position – in other words, that the probability mass was distributed over many possible words rather than concentrated in just one or two. Thus, if our measures of prediction error apply generally across all contexts, we should observe strong main effects of predicted word length and prediction error, with only weak or statistically insignificant interactions with entropy. On the other hand, if we observe strong interactions between entropy and our prediction measures, with no main effects, then the results would suggest that our measures of prediction are related to the EEG signal only at certain (most likely high) levels of predictability.

4.3 EEG data analysis

The target of analysis in study 3 was participants’ EEG responses to all non-target words contained in the sentence stimuli from studies 1 and 2. As described in previous chapters, each sentence stimulus contained one target word of interest, whose properties were experimentally manipulated to address the hypotheses of studies 1 and 2. In contrast, the other words that served as the target word’s
sentence context varied more-or-less randomly, and were chosen only to provide predictive constraint for the target word. Nevertheless, using the model-based measures outlined above, it is possible to obtain estimates of form-level prediction for these non-target words as well, and use them to predict the EEG response generated while the non-target words were read by participants.

Because studies 1 and 2 used similar sentence materials, and employed the same SOA (stimulus timing), we combined them into one large dataset for Study 3. This amounted to 173541 epochs of EEG data (or equivalently, the same number of word instances) from 56 subjects.

For each non-target stimulus word, epochs of EEG data spanning 0 ms (time of word onset) to 300 ms after word onset were extracted. At our sampling rate of 200 Hz, this amounted to 61 samples of EEG data per word at each of 64 channels. Each sample’s voltage was then used as the dependent variable in a linear mixed effects model with the predictors outlined above.

At each 5 msec. sample, we computed two mixed effects regression models containing the following fixed effects of interest: (1) the length of word that was actually displayed; (2) the model’s prediction of what the word’s length should be; (3) either prediction error measure 1 or 2; and (4) & (5), the interaction between entropy and (2) and (3) (the main effect of entropy was also included implicitly). The only difference between the two models was whether prediction error measure 1 or two was included. Random effects for all models were defined for participant and word identity. For computational reasons only random intercepts were fit for each random effect.

The use of 3904 unique statistical tests (61 samples * 64 channels) substantially increases the probability of false positives. To address this, our approach was to attempt to replicate the most interesting results in a held-out EEG dataset collected from different subjects and using different stimuli. The details of this dataset are described below. Testing on the held-out dataset was done only after the analysis of the EEG data from studies 1 and 2 was completed.
4.3.1 Replication on Held-Out EEG Dataset

A dataset obtained from another experiment in our lab (using the same equipment) was used to validate the effects we observed of the model-derived predictor variables on the EEG data from studies 1 and 2. The experiment used to collect the held-out dataset also involved word-by-word sentence reading, at a similar pace to studies 1 and 2 and with the same font (courier new). The dataset included EEG data from 24 participants. All EEG pre-processing steps (down-sampling, filtering, referencing, and epoching) were identical to those used in studies 1 and 2.

The stimuli used during the collection of the held out dataset differed from studies 1 and 2 in several ways. First, all sentence stimuli contained a relative clause (e.g. “the actor that the producer praised was married to a famous actress”) that was designed to test a hypothesis unrelated to studies 1 and 2. As these sentences were (on average) more syntactically complex than the sentences used in studies 1 and 2, their comprehension was likely to be more difficult. This makes the hold-out set a particularly strong test of any effects obtained from the analysis of studies 1 and 2. Finally, participants were not explicitly encouraged to predict upcoming words during the collection of the held out dataset, so any resulting effects of prediction would be difficult to explain by appealing to task demands.

4.4 Study 3 Results

4.4.1 Analysis of EEG data related to Study 1 and 2 Non-target words.

We report our analysis of non-target word data from studies 1 and 2 in two separate ways. First, we take a more focused approach by computing an analogous set of analyses to those used in Study 2, to make comparisons to those effects more straightforward. Namely, we split the time-window spanning 75 to 225 msec. post-stimulus into intervals of 25 msec. apiece and computed regression models on each interval’s mean voltage. Second, we took a broader approach and computed the same
models on every 5 msec. sample included in the time-window 0 to 300 msec. post-stimulus. This allowed us to look beyond the focused set of occipital electrodes used in Study 2 and exploit our roughly 10X increase in statistical power to look for effects of prediction and prediction error in other areas of the scalp montage. Below we report the more focused analyses first, and then describe important patterns that emerged in the broader analysis.

**Focused Analyses of Occipital Electrodes, 75-225 msec.**

Only t-statistics for slopes are reported below due to the computational time needed to calculate slope p-values for such a large number of mixed-effects models on a dataset of this size. The critical t-value for statistical significance at alpha = .05 (two-tailed) would be approximately +/- 1.96; alpha = .01, +/- 2.58; alpha = .001, +/- 3.30.

At occipital electrodes from 75-225 msec., we found that the relationship between (model-based) predicted word length and voltage was significant at only one time interval (175-200 msec; \( \beta = -.093, t = -2.52 \)). Furthermore, we did not find any reliable interactions between entropy and either predicted word length or the prediction error measures (all t’s < |1.85|). In what follows we therefore focus only on the main effects of word length and prediction error.

From 75-100 msec. post stimulus onset we found no relationship between changes in voltage and either word length or prediction error (all p’s > .24).

From 100-125 msec. we found that voltage became more positive as the length of the word increased (\( \beta = .17, t = 6.96 \)), and became more negative as prediction error measure 1 increased (\( \beta = -.065, t = -2.20 \)). See figures 3.3 and 3.4a.

From 125-150 msec. we found that voltage became weakly more positive as the length of the word increased (\( \beta = .05, t = 1.95 \)). We found no effect of either prediction error measures (t = .57).
For the rest of the three time intervals within the window 150-225 msec. post-stimulus, we found that voltage became more negative as the length of the word increased (mean $\beta = -0.37$, mean $t = -15.16$), and became more positive as both measures of prediction error increased (measure 1: mean $\beta = 0.20$, mean $t = 6.92$; measure 2: mean $\beta = 0.05$, mean $t = 2.71$).

Figure 3.3 Right: ERP effect of word length for all non-target words in studies 1 and 2. Data are split into quartiles for display purposes only. The data show a clear increase in N170 amplitude as the character length of the replacement string increases. Left: scalp distribution of t-statistics associated with the word length parameter estimate (slope), when predicted word length and error size are controlled for. The scalp locations used for the focused analysis of 75-225 msec. are outlined in black. Each scalp plot is the average over 25 msec. of time, starting at 75 msec. (top left) and ending at 225 msec. (bottom right). Bottom right: scatter plot of relationship between word length and voltage, 175-250 msec., when controlling for error size and predicted word length. On the y-axis are the residuals from a model including predicted word length and error size, on the x-axis is word length.
Figure 3.4a Right and top: ERP effect of Measure 1 error size for all non-target words in studies 1 and 2. Data are split into quartiles for display purposes only. The data again show a small but reliable increase in positivity with error size. Left: scalp distribution of t-statistics associated with the error size parameter estimate (slope), when predicted word length and actual word length are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 75 msec. (top left) and ending at 225 msec. (bottom right). Bottom right: scatter plot of relationship between error size and voltage, 150-250 msec., when controlling for word length and predicted word length. On the y-axis are the residuals from a model including predicted word length and word length, on the x-axis is error size.
Figure 3.4b Right: ERP effect of Measure 2 error size for all non-target words in studies 1 and 2. Data are split into quartiles, numbered from 1 (smallest errors) to 4 (largest errors). The grey box indicates the time-window in which the linear effect of word length is significant at $p = .05$. The data again show an increasing positivity with error size. Top left: scalp distribution of t-statistics associated with the error size parameter estimate (slope), when predicted word length and actual word length are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 0 msec. (top left) and ending at 300 msec. (bottom right). Bottom right: scatter plot of relationship between error size and voltage, 175-225 msec., when controlling for word length and predicted word length. On the y-axis are the residuals from a model including predicted word length and word length, on the x-axis is error size.
Figure 3.5: Right: ERP effect of predicted word length for all non-target words in studies 1 and 2, center parietal channels. Data are split into quartiles for display purposes only. The data show an increasing central-parietal positivity with longer predictions. Left: scalp distribution of t-statistics associated with predicted word length parameter estimate (slope), when error size and actual word length are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 75 msec. (top left) and ending at 300 msec. (bottom right). Bottom right: scatter plot of relationship between predicted word length and voltage, 75-150 msec., when controlling for word length and error size. On the y-axis are the residuals from a model including error size and word length, on the x-axis is predicted word length.

**Broader Analyses of all Electrodes, 0-300 msec.**

We undertook a broader analysis of the scalp EEG by analyzing each electrode separately, using the same regression analysis described above. We visualized these results by mapping the t-values
associated with the slopes for word length, predicted word length and prediction error, onto the scalp, as shown in figures 3.3-3.5. We discuss the effects of each predictor in turn below.

Beginning with word length (figure 3.3), we find that the relationship with the EEG signal follows the well-known pattern of larger P1 amplitudes (approx. 100-150 msec.) and larger (i.e., more negative) N170 amplitudes (approx. 150-250 msec.) as words contain increasing numbers of characters. These effects appear most robust (t-values ranging from 10-20) in expected occipital scalp locations. It also appears that there may be a very early (50-100 msec.) effect of length at right-temporal/center-parietal scalp locations, though I am not aware of any prior work reporting EEG effects of stimulus size at this latency. The effect’s dipolar pattern is suggestive of true brain activity, however.

Turning to predicted word length (figure 3.5), the principle effect appears to be an early positive correlation with the EEG at center-parietal electrodes. The effect begins immediately after stimulus onset and remains roughly constant through 150 msec (peak t-value before 150 msec. = 5.10, 70 msec. at channel PZ). The effect become stronger after about 200 msec. Due to its scalp location and morphology it is not entirely clear that this truly represents a prediction of visual length (especially at later time points), or instead another variable correlated with either the current or previous word’s length.

Considering now the effects of prediction error measure 1 (figure 3.4a), these are largely consistent with those observed in Study 2, where prediction error was manipulated explicitly. The main difference is that the effect begins slightly earlier (approx. 125 msec.), at least at left posterior scalp sites. The effect is strongest in the interval of 150-200 msec. (peak t = 10.85, channel CB1 at 180 msec.) corresponding to the rising slope of the N170 ERP. The effect has a dipolar pattern largely consistent with a generator in visual cortex. This potential source is further supported by the fact that the
correlation between the EEG and error size is more reliable at the posterior rather than anterior half of the dipolar pattern.

For prediction error measure 2, we found no notable correlations outside those reported in the more focused analysis above (see figure 3.4b).

4.4.2 Replication of Findings in Held-out Data

To ensure the generalizability of our findings, we attempted to replicate the most substantial of the effects from Study 3 in a focused analysis of another dataset. As described above, the data were collected from different participants reading different sets of stimuli in an unrelated experiment. We focused on three effects from our analysis of the non-target words from studies 1 and 2. The first was the basic effect of word length from 100 to 225 msec (figure 3.3) at occipital electrodes. The second was the effect of predicted word length from 0 to 150 msec. at center-parietal electrodes (figure 3.5). The third effect of interest was our effect of prediction error measure 1 from 150-225 msec. at occipital electrodes (figure 3.4a). Our analysis employed the same type of regression model as that is reported above. For the occipital effects we used the same electrode group (OZ/1/2, POZ/7/8, CB1/2). For the central parietal effect of predicted word length we defined an electrode group centered on the electrode PZ, where the effect was strongest (PZ/1/2, CPZ, POZ). Because the effect of word length 75-225 msec. changes direction as the P1 transitions to the N170, we tested two time-windows 100-150 msec. and 150-225 msec. This amounted to three separate regression analyses: 0-150 msec. at center-parietal electrodes (concerning the effect of predicted word length), 75-150 msec. at occipital electrodes (concerning the effect of word length on the P1), and 150-225 msec. (concerning the effect of prediction error). Below we report the results of each of these in turn.

Regarding our analysis of center-parietal electrodes, we failed to replicate our early effect of predicted word length from 0-150 msec. (β = -.05, t = -1.28, p = .18). From 100-150 msec., we found
some marginal evidence for a positive correlation between the word length of the displayed stimulus and P1 amplitude ($\beta = .11, t = -2.05, p = .049$), but stronger evidence for the relationship between N170 amplitude and word length ($\beta = -.39, t = -3.67, p < .001$; see figure 3.6). Finally, we were able to replicate the positive correlation between prediction error and voltage from 150-225 msec. ($\beta = .19, t = 3.14, p = .003$, see figure 3.7).

Figure 3.6: Top left: ERP effect of word length for the held-out data in study 3. Data are split into quartiles for display purposes only. The data replicate the increase in N170 amplitude found in the analysis of study 1 and 2’s non-target words. Bottom: scalp distribution of t-statistics associated with the word length parameter estimate (slope), when predicted word length and error size are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 100 msec. (left) and ending at 300 msec. (right). These t-maps were calculated for display purposes only, to make them more easily comparable to the previous results. Center right: scatter plot of relationship between word length and voltage, 150-225 msec., when controlling for error size and predicted word length. On the y-axis are the residuals from a model including predicted word length and error size, on the x-axis is word length. Each point is an average over each stimulus word.
Figure 3.7: Top left: ERP effect of error size for the held-out data in study 3. Data are split into quartiles for display purposes only. The data replicate the positive relationship between error size and voltage found in analysis of study 1 and 2's non-target words. Bottom: scalp distribution of t-statistics associated with the word length parameter estimate (slope), when predicted word length and error size are controlled for. Each scalp plot is the average over 25 msec. of time, starting at 150 msec. (left) and ending at 225 msec. (right). These t-maps were calculated for display purposes only, to make them more easily comparable to the previous results. Center right: scatter plot of relationship between error size and voltage, 150-225 msec., when controlling for word length and predicted word length. On the y-axis are the residuals from a model including predicted word length and predicted word length, on the x-axis is error size. Each point is an average over each stimulus word.

In sum, the effects we were able to replicate from our analysis of the non-critical words in studies 1 and 2 include the effect of prediction error on the P1-N170 transition, as well as the effects of
word length on the P1 and N170 (though the relationship was clearest for the N170). We were not, however, able to replicate the effect of predicted word length at center-parietal sites.

4.6 Study 3 Summary & Discussion

Study 3 extended the results of studies 1 and 2 by showing that form-level predictions during reading occur in contexts that do not necessarily support a firm commitment to the particular identity of the next word. This was established by using a computational model to estimate a probability distribution over the identity of each non-target word in the stimulus sentences used in studies 1 and 2, and leveraging this distribution to generate a prediction about what the lengths of each of these words “should be,” given the sentence context. Our principle finding was that the deviation of this prediction from the actual length of each word was related to the early visual EEG response approximately 150-225 msec. post-stimulus onset. In particular, we found that larger such deviations (whether they be under- or over-estimates of the word’s length) lead to more positive voltages during the later phase of the P1 ERP component, and during the early part of the N170. This effect was in the same direction, at the same occipital electrodes, and at approximately the same time-window as the pattern we observed in Study 2’s explicit manipulation of prediction error. Importantly, we were able to further replicate this effect in a held-out EEG dataset obtained from a separate sample of participants who read different stimulus sentences.

We also observed an effect consistent with form prediction, rather than prediction error, though this effect did not replicate in the held-out data, and also differed from the reduction of P1 amplitude found in Study 2. The effect was characterized by a concomitant increase in center-parietal positivity and model-derived predicted word length. Beginning immediately at stimulus onset, the effect continued throughout the initial 300 msec. of the brain response.
Taken together with the analogous effect from Study 2, Study 3’s effect of prediction error strongly suggests a role for form-level prediction during visual word recognition. As will be discussed in more detail below, this finding has the potential to bridge a gap (of sorts) between the processes underlying word recognition and broader theories of perception that posit a central role for prediction. While the effects attributable to “bottom-up” processing (i.e. our effects of word length) were considerably stronger than our reflects of prediction, the clear influence of prediction error, quantified in two different ways and in three separate analyses, represents a neglected source of variance in most models of reading and comprehension.

Finally, the results of Study 3 show how advances in computational linguistics and machine learning can be leveraged to further the cognitive neuroscience of language, especially when EEG is used as a dependent measure. Neural network models of language sequences have now advanced to the point that they can (provided they are trained on the appropriate corpus) be used as reasonable proxies for the distributional knowledge of most humans. Indeed, it is possible to train much more accurate models than the one reported here, provided one has the computational resources and time. As I will discuss in more depth in the conclusion, these models can be used to provide incremental estimates of the transitional probabilities involved at all levels of linguistic representation (e.g., one could estimate the probability of a phoneme given previous phonemes). Though these tools cannot be used as direct models of the psycholinguistic processes themselves, they nevertheless hold considerable promise as a method for estimating the “objective” probabilities associated with language productions, and therefore the amount of surprise or confusion that should arise when they are processed by human language comprehenders.

5. General Discussion
Over the course of three studies we observed repeated evidence that language comprehenders make predictions about the low-level forms of upcoming words. In the first study, we found that P1 amplitudes in response to an omitted word still reflected the expected word’s form characteristics. Study 2 then manipulated prediction error by replacing predictable words with a variable-length symbol string, rather than an empty display prompt. This allowed us to disentangle effects of prediction error from prediction generation, and appeared to confirm that the former precedes the latter in time. Finally, Study 3 aimed to establish whether form-level predictions occur outside of highly predictable sentence contexts, by using a computational model to “simulate” what form-level predictions may have been generated by our participants at every (non-target) word used in studies 1 and 2. We found that our model-derived measure of prediction error (analogous to the one used in Study 2) did in fact correlate with the amplitude of the P1/N170. This effect of prediction error also held up in an analysis of a separate dataset.

Taken together, these results have important consequences for the functions underlying word recognition and language comprehension, and may represent a bridge to general predictive coding models of perception and cognition.

With regard to word recognition in particular, our findings suggest that the brain’s early visual response to a word represents a complex combination of top-down prediction, bottom-up response, and error detection. This is consistent with the fact that the early EEG response to a visual stimulus reflects the mixture of many different neural sources (Makeig et al., 2002). Our results therefore represent a step towards a more biologically plausible description of the processes underlying these early visual responses, and go beyond earlier models that suggest a simple one-to-one mapping of functional processes to ERP components (Grainger & Holcomb, 2009). We further unpack the psycholinguistic implications of our results in sections 5.1-5.3.
Our results also have implications that go beyond language comprehension. They represent one of the strongest empirical tests (thus far) of the idea that prediction plays a fundamental role in neural processing. Incremental language comprehension is a cognitively complex process that relies on a deep hierarchy of representations; despite this, our results strongly suggest that predictions penetrate this hierarchy to its lowest levels. Thus our results represent an important step towards confirming that predictive coding models can account for specific, ecologically relevant functions like comprehension and word recognition. We return to this point in more detail in section 5.4.

The rest of the discussion section is split into five parts. The first subsection section considers the relationship of our results to previous accounts of prediction in language comprehension, and specifically addresses the issue of whether they specifically support the *pre-activation* of certain representations, which is sometimes considered to be more controversial than mere prediction. The second subsection considers the issue, raised by Jackendoff (2009) and others, of what utility prediction might have for the core functions of comprehension. In section 5.3 I consider the representational issues raised by the existence of probabilistic word-form prediction, and offer an analogy to help conceptualize what it could mean that aspects of many word-forms can be predicted in parallel. Section 5.4 then addresses the specific implications our results have for the study of the early EEG response to a visual word-form, and relates our findings to previous ERP studies of word recognition. Finally, subsection 5.5 proposes a predictive-coding based framework for word recognition in context. A concluding section offers some remarks on possible future directions.

5.1 *Prediction vs. Pre-activation*

Some prior literature has drawn subtle distinctions between what might be termed different types of Prediction writ-large, as dissociations have been made between anticipation and prediction (Van Petten & Luka, 2012), prediction and pre-activation (Kuperberg & Jaeger 2016) and even predictive
pre-activation and predictive commitment (Lau, Holcomb & Kuperberg, 2013). For our purposes, the important distinction lies between the more minimal concept of prediction and the more active concept of pre-activation, sometimes referred to as predictive pre-activation. In order for a language system to be predictive, it need only be that it changes its current state in such a way that future input be more easily handled. This might mean as little as moving one’s head towards a new source of noise. This is distinguished from pre-activation, in which the specific representations of expected future input are accessed prior to their anticipated arrival through sensory channels.

With regards to the current data, we believe that most of our effects require that some variety of pre-activation occur. In all cases, our findings related the specific – but largely arbitrary – feature of predicted word length to the amplitude of the occipital visual EEG. In order for such a relationship to exist, there must have been some specific aspect of the visual word form that was activated prior to when the correlation was observed. Put another way, there is no principled reason why the signal would reflect the incorrect, never-actually-displayed word length (as in our findings regarding prediction error) after the “true” length of stimulus was revealed by bottom up input. The existence of a discrepancy between two potential word lengths (one expected and one actually observed) necessarily implies an anticipatory pre-activation of some expected form. The case becomes even stronger when we consider the results of Study 1, in which brain responses to an empty display still reflected the lengths of the expected (but never encountered) word.

In this respect our evidence for pre-activation is somewhat less open to alternative explanation than the majority of EEG results that have been argued to represent a similar process. In particular, pre-activation has been an explanation for many findings involving changes in N400 amplitude in response to manipulations of semantic predictability (Federmeier & Kutas, 1999; Delong, Urbach & Kutas, 2005). It has been pointed out, however, that changes in N400 amplitude can instead be explained by difficult
(in the case of unpredictable words) or facilitated (in the case of highly predictable words) semantic integration, in which the presence of related words in the prior context eases the retrieval of the meanings associated with upcoming words (Federmeier 2007). Importantly, such an account would be Predictive in the minimal sense of the term, but without necessarily invoking pre-activation. For instance, it is easy imagine that facilitated access of a semantically similar word can happen “incidentally” due to many of the word’s semantic features being “residually active” as a consequence of processing previous, semantically related words. This is different from a case in which the features of the new word, including those that were not shared by the previous words, are specifically targeted for activation prior to word onset. Because the feature of word length is largely arbitrary, (words sharing semantic features don’t systematically have similar lengths), our results do not seem explainable by a passively predictive mechanism like those sometimes posited to explain N400 effects.

5.2 The utility of low-level form predictions

Language comprehension is a complex function that relies on a complex web of dependencies between words that is developed over a lifetime of experience. These dependencies provide affordances for prediction by constraining the probability distribution over upcoming words in a discourse. However, some have argued that these statistical patterns are only a by-product of the more fundamental grammatical and semantic relationships that are ultimately used to compute the meaning of a sentence (Jackendoff, 2007). If these statistical regularities were purely epiphenomenal, however, then we would not expect them to be exploited for prediction, particularly for predictions of word forms. Given that our results do suggest pervasive low-level prediction, we must conclude some combination of either (1) that such predictions directly aid in comprehension, or (2) that the neural machinery underlying word recognition is itself predictive. Although our results do not directly
distinguish between (1) and (2), there are independent arguments for (1), such as the fact that prediction may free up resources (attention, for example) that are directed toward drawing inferences from the rapidly arriving input. However, accepting only that (2) is true represents the more conservative conclusion, and in doing so one might argue that low-level prediction confers only an *indirect* benefit to comprehension, incurred as a result of its dependence on the broader perceptual system. I do not find such a distinction that useful; clearly, the language system relies on the more general architectural principles of the brain, and to fully describe its function within that context one must (at some point) explain how these more general principles allow for the specific functions of language.

It can also be argued that, while every sentence context contains *some* information about likely upcoming language, very often this information will not be enough to support the prediction of a particular word. This appears to conform to our intuition that, if we had to predict one word to continue each sentence position, then we would often be wrong. Given this state of affairs, one might conclude that only certain types of predictions will be useful in the average context. Specifically, predictions of high-level features that are compatible with many possible inputs would appear to be a “safer” bet than predictions that represent commitments to arbitrary features – such as the visual form – of particular words. Nevertheless, the results of the studies reported here provide strong evidence that such low-level predictions do in fact occur. Study 3 in particular shows that such predictions may not always be of one particular word’s form, implying that form predictions are implemented probabilistically as a way of hedging against language’s inherent uncertainty. Below I speculate on what the probabilistic nature of form prediction may suggest about the representation of word-forms more generally.

5.3 *The representational nature of form predictions*
Our measure of prediction error in Study 3 was defined at every stimulus word by taking a sum of the lengths of each word in our model’s vocabulary, weighted by each word’s probability that it occur at that position. Representationally, this implies a graded “superposition” of many different word forms. Representing predictions as a weighted combination of many possible outcomes, rather than commitment to a single one, has a clear advantage in terms of efficiency. As mentioned, cases where the outcome is underdetermined still often contain cues, such as the parts of speech of prior words, which can provide some information about the form of upcoming language. Indeed, analyses have shown that nouns are closer to other nouns, in terms of their phonology, than they are to verbs (even when verb morphology is taken into account; Monaghan, Christiansen, Farmer & Fitneva, 2010). To ensure that such a claim extends to the visual forms of words, I have made a similar analysis of the orthographic forms of the most frequent 4,000 English nouns and verbs, which is depicted in figure D1. This shows that, on average, it takes fewer edits (insertions, deletions or transpositions) to turn a noun into another noun, than it does to turn a noun into a verb (and similarly for verbs, they are more similar to one another than to nouns). This suggests that, given a prior context such as “the large...”, a prediction of something like the average noun-form will usually be a closer match to the eventual word than no prediction, or a default prediction of something like the language-average word-form. The weighted measure used in Study 3 often resembled something similar to a part of speech-based prediction at earlier sentence positions, where prior context did not provide much more than grammatical information with which to make a prediction (many sentences simply start “The...”). Given that the deviation of the actually displayed word from this weighted measure was predictive of the EEG even at these early sentence positions, our results strongly suggest that language comprehenders exploit even the subtlest form of linguistic regularities to make form-level predictions during reading.
Orthographic Similarity of Nouns and Verbs

Figure D1: Analysis showing the average orthographic edit distance (number of insertions, transpositions or deletions needed to change one written word into another) of the 4000 most common English nouns (filled circles) to all other nouns and verbs, as well as the average edit distance of the 4000 most common English verbs (triangles) to all other nouns and verbs. Both plots are the same, the only difference is whether nouns or verbs are plotted on top. The analysis shows that most nouns are more orthographically similar to nouns than verbs (circles falling above the diagonal) and that most verbs are more similar to other verbs (triangles falling below the diagonal). This suggests that there may be some aspects of written form that are predictable solely based on word category.

One way to conceptualize the representations of word forms during prediction would be to make an analogy to vector algebra. If individual word forms are represented in a distributed fashion (as they almost certainly must be, given what we know about neural coding) then we could define each unique word-form as a particular linear combination of some finite set of basis vectors. Not all (or even most of the) combinations of the basis set might correspond to a particular word – instead, certain
combinations may instead represent a complex combination of many forms, similar to the superpositions discussed above. If word forms were to be encoded in such a way, then we would not be surprised that uncategorizable, weighted combinations of features are mainly what are computed during word-form prediction. Making a form-level prediction would then simply involve the right combination of the basis set, such that the resulting “prediction vector” points in a direction that is optimally close to the most likely upcoming word-forms. Furthermore, if the basis set is carefully chosen, then it may be that the encoding of word-forms can be made much more efficient than in a system where each form is represented independently.

Our results therefore constrain the nature of how word-forms must be represented, albeit in an indirect way. Because form predictions appear to occur in less-than-constraining contexts, it suggests a system which is capable of efficiently computing “in-between” form representations that are somehow superpositions of many different orthographic features. A good analogy for such a scheme is a vector space in which certain combinations of features represent actual words, while others represent a weighted combination of many. Ensuring that a “predicted” vector is appropriately close to the most probable next words in the sentence will (on average) reduce the work the system must do to accurately represent the incoming sensory information.

One criticism of such a representational scheme may be that the “in-between” combinations of basis vectors that occur during prediction are not fundamentally distinguished from the “true” representations of actual word-forms; if any combination of the basis vectors is (in theory) permissible, then there exists no categorical boundary between combinations that represent actual word forms from those that represent in-between combinations. Clearly, there must be some aspect of a word’s form representation that distinguishes it as an actual word, rather than an uninterpretable combination of perceptual features. I would argue, however, that the categorical aspects of form representations are
due to the fact that particular combinations of orthographic features are linked, at higher levels, to more abstract semantic features. Here again, higher level representations provide constraints on the combination of certain lower-level representations, such that only certain combinations of perceptual features are “linked” to their associated semantic representations at higher levels. Crucially, this does not prevent the “meaningless” combination of features to occur during form prediction, as a consequence of the fact that general, lexically non-specific semantic or grammatical features can be associated with a wide variety of lower-level form features.

5.4 The relationship of early visual ERP components to word recognition

In studies 2 and 3, we examined the timing of effects related to form-level prediction, prediction error and bottom-up processing with a finer temporal resolution than is typically used. Doing so lead to a complex picture of how the early visual EEG is modulated by both predictive and bottom-up factors, suggesting that early visual ERP components are not purely exogenous in nature, as some previous accounts have assumed.

The contribution of predictive processes to the morphology of the P1 and N170 are not well represented in existing accounts of their underlying functions. For example, Grainger & Holcomb (2009) describe a framework for orthographic and phonological word recognition that explicitly maps stages of visual word processing onto separate peaks of the ERP, including the P1 and N170. Importantly, they base their model primarily on results from masked priming experiments using single word presentation, and thus do not consider the possibility of context-based prediction effects. Despite this, their architecture stands as one of the more influential attempts to define a clear mapping of electrophysiology to the stages of word recognition, and thus it is important to address how, in light of the current results, their model must be revised if it is to account for word recognition in context.
In their Bi-Modal Interactive Activation Model, Grainger & Holcomb (2009) define a clear progression of form analysis that begins with low-level visual features (oriented lines and the like), moves onto analysis of individual orthographic features, to complete visual word-forms and finally to semantic representations. These stages are then mapped onto the P/N150 (corresponding to our P1, and the anterior, negative side of its dipole), the N250 (corresponding to our N170), the P325 and N400, respectively. Grainger & Holcomb (2009)’s model is grounded in the hierarchical nature of the ventral visual stream (Milner & Goodale, 1992), which posits stages of feature detectors with progressively larger and more complex receptive fields. The model is designed with bi-directional connections between levels of analysis, such that (for example) semantic representations can influence processing at lower, sub-lexical form levels. Critically, however, these connections do not serve the purpose of conveying predictions and prediction error, as in a predictive coding model. Rather, they are more closely related to the sorts of connections found in early connectionist models (e.g. McClelland & Rumelhart, 1981). These top-down connections instead serve to strengthen or “clean up” activation at lower levels, due to the ability of higher levels to become partially active before lower levels completely stabilize their activation patterns.

The functional difference between the feedback connections present in “classic” connectionist models like Grainger & Holcomb (2009) and those used in predictive coding models (e.g. Rao & Ballard, 1997) represent differing theoretical points of view on neural information processing. In standard connectionist models of word recognition, processing at the current time-step (i.e., word) always begins afresh at the bottom of the processing hierarchy. Even in most recurrent models of language processing (e.g. Elman, 1991), information about previous time-steps is incorporated as an additional feed-forward input. Feedback connections then work to stabilize activity at lower levels, rather than actively generate ‘hypotheses’ about what the lower-level activity should be, as in a predictive coding architecture.
Furthermore, consider a comparison between recordings obtained separately from a standard connectionist model and a predictive coding architecture. In the former case, activity observed in response to a word would simply reflect the activation of the word’s semantic and perceptual features, and the communication between those two levels of representation. In the case of the predictive coding architecture, the data would instead represent some combination of externally-driven responses (in the form of prediction error) and internally generated, top-down predictions about what the externally-driven response should be. This is more reflective of the results of studies 2 and 3, in which the purportedly “exogenous” P1 and N170 components in fact reflected a combination of what the form of the current word actually was, and what it was most likely to be, given the sentential context.

Thus, it appears that “stage of processing” is only one factor that contributes to the early brain response following a word. According to the current results, we must also consider top-down hypotheses that brain generates about what linguistic content should be at any point in time. This somewhat complicates the job of psycholinguistic researchers who often think of ERP components as truly “event related”, in the sense that brain activity time-locked to the onset of a linguistic stimulus mainly reflects the integration of whatever new information the word represents. Instead, it appears that even very early, sensory-specific signals are not immune from the influence of top-down expectations. This in turn implies the need for researchers to more carefully separate out aspects of the brain signal that truly reflect sensory responses to a linguistic stimulus, from those that instead reflect internally generated hypotheses about what those sensory responses should be, given prior knowledge and input.

5.5 A predictive coding account of word recognition?

In Study 2, we found evidence that the brain’s response to a word may reflect a sequence of prediction generation, bottom-up processing, and the computation of prediction error. Rather than each
of these processes mapping neatly onto the peaks of particular ERP components, we found that prediction-related activity was related to the earlier stages of the P1 (though these effects were not as reliable), while the later stages showed correlations between stimulus length and prediction error. Indeed, the ordering of results from Study 2 are strikingly reminiscent of predictive coding models in which higher levels of representation send predictions to lower levels, which in turn compute the error between these predictions and bottom-up input.

Study 1 showed that early visual brain responses to predictable – but omitted – sentence embedded words still reflected aspects of the expected word form. Considered in isolation, this correlation could have reflected either the generation of a prediction related to word-length, or instead the prediction error between the length that was expected and the crosshair prompt that replaced the word. To follow this finding, studies 2 and 3 showed that prediction error made a portion of the P1 more positive, though in some cases the correlation was only reliable on what would be considered the later half of the P1, and the beginning of the N170 (though still overlapping with the time-window that we used analyze the results of Study 1). Furthermore, in Study 2 we found that the length of highly predictable target words that were replaced by non-word character strings showed a negative correlation with P1 amplitude, opposite of what we would expect had the pattern of Study 1 reflected predicted word length rather than prediction error (the statistical support for this effect was considerably weaker, however). On balance, then, the direction of our prediction error effects, and to a lesser extent their timing, appears to support the notion that the primary result of Study 1 in fact reflects prediction error, rather than predicted word length.

The empirical evidence thus supports an early (around 150 msec. post-stimulus) positive effect of prediction error, while effects of prediction itself on visual components appear to be negative (c.f. Study 2) and less robust. On the surface, our overall pattern of results therefore appears compatible
with predictive coding models of perception and cognition (Rao & Ballard, 1999; Friston & Kiebel, 2009; Clark, 2013). Friston & Kiebel (2009) even describe a predictive coding-based simulation of birdsong perception in which omissions of part of a song sequence lead to larger prediction errors during recognition. Such results can in turn be taken as a model of studies on the so-called auditory mismatch negativity (MMN) ERP that occurs when a predictable sequence of simple tones unexpectedly deviates in some way, for example when there is a random omission (Nordby, Hammerborg, Roth & Hugdahl, 1994; Yabe, Tervaniemi, Reinikainen & Näätänen, 1997). The fact that such omissions do lead to MMNs (at least at short SOAs, 100-120 msec.) is difficult to explain without invoking some notion of either prediction or prediction error. Although our results concerned a different component of the ERP, the overall situation is very similar, and in one sense even more indicative of prediction error than the MMN studies. This is because the magnitude of the “error” in our case actually varied from word to word (at least in studies 2 and 3), and the EEG signal seemed to reflect this.

It is worth an attempt at a more detailed mapping of our results onto specific predictive coding architectures, in order to see just how well they might account for our results. In both the models of Rao & Ballard (1999) and Friston & Kiebel (2009), the only messages that are passed between levels of the neural hierarchy are the predictions themselves (via top-down connections) and prediction errors (via feedforward connections). The predictions themselves represent the higher level’s best hypothesis about what the lower level’s activity should be. The lower level then compares its current state to these predictions, passing any discrepancy back up the hierarchy as prediction error, which in turn is used to by the higher level to appropriately adjust its hypotheses. Each layer of the hierarchy maintains two separate but interacting populations of neurons that work together to accomplish these functions. The first class of “representation” units (R-units) encodes the layer’s current estimate of the input, as a probability distribution over all its possible states, and forwards this estimate to the layer immediately below (in the cortex, backward connections originate in pyramidal cells found in the infragranular layer;
The second class of unit (E-units) calculates the discrepancy between the state of the R-units and the predictions received from layer immediately above, and forwards the resulting error back up the hierarchy (forward connections arise in pyramidal cells found in supragranular layers; Felleman & Van Essen, 1991). This arrangement minimizes the system’s long-term prediction error and entropy, and leads to an efficient encoding of the input.

In figure D2 I’ve sketched out a cartoon of how the posterior occipital-temporal language system might implement a predictive coding model that is consistent with our results. It assumes a hierarchically organized taxonomy of representations with sub-orthographic features (e.g. oriented edges) at the bottom, and more abstract representations that encode semantic and/or syntactic categories at the top. I’ve drawn the state of the hierarchy at two moments in time after the input of a new word. At the left, predictions about forthcoming input are imposed at lower levels, based on prior knowledge and contextual constraints (assumed to be encoded largely by the highest levels in the hierarchy). We assume that at (or slightly before) 150 msec., information about the new stimulus enters the hierarchy from the thalamus (based on the earliest time-windows in which we observe reliable correlations with word length). At this point, prediction error is calculated as the difference between predictions made about the word’s expected form prior to thalamic input, and the new information, beginning at the lowest levels of the visual ventral stream and cascading forward. This stage is represented in our data by our prediction error effects, observed approximately 150-200 msec. post-stimulus. R-units are then able to change their activity in response to the prediction error, and settle on a stable representation of the current percept, which in turn allows the system to generate new predictions about what future states should represent.
Figure D2: Sketch of a predictive coding framework for word recognition in context. The diagram is split into two time intervals, before 100 msec. (left) and 150-200 msec. (right). Blue boxes represent so-called representational units (R-units) that represent probability distributions (black lines in blue boxes) over the current representation of the stimulus. Red boxes represent error units (E-units) that calculate the discrepancy between top-down predictions and the layer’s current estimate of the input. The top boxes are intended to represent high-level lexicosemantic representations, and the bottom boxes are intended to represent pre-orthographic from units. Dotted lines represent top-down (predictive) connections, solid curved lines represent feedforward error signal propagation. Solid horizontal lines represent bi-directional, lateral connections between R and E-units. Connections in red represent those most active at each time interval.

This framework captures the most essential aspects of our data, most importantly the fact that effects of bottom-up stimulus features (e.g., length) begin around the same time (or slightly before) our effects of prediction error, which we infer incorporates the system’s hypothesis of what the input should have been, prior to the word’s actual onset. Furthermore, the strongest effects of word length occur
after this time window, suggesting that the brain’s representation of the current word is at its most veridical once prediction error has had a chance to have an impact.

Finally, we must be clear that while we believe this framework to be a step towards a more dynamic model of word recognition in context, there is still much more to be done in terms of identifying its specifics. First, there is still a considerable lack of specificity when it comes to mapping identifiable stages of word recognition onto subdivisions of the occipital-temporal ventral stream. Developing a more specific functional mapping would allow us to better test this framework, as it would constrain the exact nature of the prediction error that is computed at each level of the system. There are many ways in which a word can deviate from expectations – with regard to form, meaning and word category, to name the obvious. If we knew what stages of the ventral stream roughly corresponded to each of these levels of representation, then mapping changes in the brain response due to error at each of these levels could be done more systematically. Second, with regard to form predictions in particular, there is still a great deal to be learned about the neural representations of orthographic features. In section 5.5 I speculated on a vector encoding scheme in which word-forms – and combinations of word-forms – might be efficiently represented. Whatever the exact format, more detail about the representational specifics of word forms will be needed in order to understand what the functional unit of prediction error is at this level (it is unlikely to be number of characters, for example).

6. Conclusions and future directions

Over the course of three studies, this dissertation has looked in-depth at the question of just how predictive the human language system actually is. A consistent effect of prediction error was observed across three separate analyses, implying that predictions play a role at even the lowest levels of form representations. As we explained in the introduction, such a result was far from a foregone conclusion. For one, it was not obvious, from an a priori perspective, that constant prediction would be
a “winning” strategy, from a computational point of view. A lack of evidence for low-level prediction
could easily have been explained by concluding that early sensory input is simply too noisy to be
predicted with better-than-chance success. Nevertheless, our results seem to suggest that prediction is
more pervasive and deeper-reaching than previously assumed.

I’d like to conclude by briefly mentioning some potential future directions. From a technical
perspective, there is considerable potential use for LSTMs (or similar models) as models of a kind of
“ground truth” about the conditional probabilities contained within a sentence or discourse. Put
another way, these approaches provide researchers with a relatively painless way to quantify the
predictability of each sequential element of a language stimulus, which in turn can be used for a number
of purposes. Importantly, predictability can be quantified at many levels of representation – phonemes,
syntactic phrases, semantic features, etc. In principle (i.e., assuming enough training data is available),
one could train a predictive model at every level of the language hierarchy, and then identify distinct
aspects of the EEG signal that is related to prediction error at each level (and whether evidence of
prediction error conveyance between levels can be more precisely specified). This could result in an
extremely detailed predictive coding model of the language system.

There are also ways to extend the experimental paradigms used here. For one, the use of
character length in the present work was intended as only an indirect measure of low-level perceptual
form. There are certainly more sophisticated ways to quantify the visual form of a word, and it is worth
thinking about what measures of perceptual form are more likely to be closer to the underlying neural
representation. One might even re-use the paradigm of Study 1 to compare competing schemas for
orthographic representation, with the one that best tracks the EEG signal during stimulus omission being
favored as closest to the “actual” underlying neural representation. Finally, one might manipulate other
features of the omitted stimulus – such as its meaning, frequency, or phonological representation – to
gain a better understanding of how prediction might operate at other levels of the language hierarchy.
References


