Probabilistic linguistic representations: between learning and processing

by

Tal Linzen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Linguistics
New York University
September, 2015

_____

Alec Marantz

Acknowledgements

First and foremost, I would like to thank my advisor, Alec Marantz, who sets a personal example of interdisciplinary research on language and cognitive science and encourages his students to pursue this type of research themselves. Alec's patience and support, as well as his insightful suggestions about any scientific topic imagineable, have been crucial in all of the work that I have done during my time at NYU, in this dissertation and outside it. I also thank my NYU committee members, Gillian Gallagher, Maria Gouskova and Liina Pylkkänen, for productive collaborations and invaluable advice, as well as my external committee member, Florian Jaeger, for hosting me at the University of Rochester and collaborating on the project that turned into Chapter 3 of this dissertation.

I'm also grateful to my fellow students at NYU. Particular names that stand out are Joe Fruchter, Itamar Kastner, Sonya Kasyanenko, Jeremy Kuhn, Gwyneth Lewis, Allison Schapp and Daniel Szeredi. I would like to single out Masha Westerlund, who apart from being a close collaborator on multiple projects was also a loyal friend and a reliable source of support through a lot of the inevitable vicissitudes of a Ph.D. program. I also thank the members of the HLP lab at the University of Rochester, in particular Judith Degen, Alex Fine and Dave Kleinschmidt.

For help with data collection, I am grateful to Rebecca Egbert, Allyson Ettinger, Phoebe Gaston, Laura Gwilliams, Andrew Watts and Jeffrey Walker. Finally, the work presented in this dissertation has benefited from discussions with a many people in addition to my committee

members and fellow students; a partial list includes Frans Adriaans, Jon Brennan, Mike Frank, Todd Gureckis, John Hale, Bruno Nicenboim, Tim O'Donnell, David Poeppel, Brian Roark and Nathaniel Smith.

Abstract

Our interactions with language often lead us to consider multiple hypotheses simultaneously. As we read a sentence, we anticipate upcoming structure; there are normally many such continuations. When we acquire a language, our input is typically compatible with many hypotheses about the rules of the language. This dissertation explores the consequences of maintaining uncertainty over representations in language comprehension and learning.

The first part of the dissertation investigates how readers' or listeners' uncertainty about linguistic structures affects comprehension difficulty. Following a theoretical discussion of uncertainty in language comprehension, three empirical studies are presented. A self-paced reading study found that words that reduce the reader's uncertainty about the syntactic structure of the sentence, as estimated from a corpus-based probabilistic grammar, can cause a slowdown in reading. An electromagnetic (MEG) study of single word recognition found that the processing of words with higher syntactic uncertainty was associated with decreased neural activity in anterior temporal regions. Finally, an MEG experiment is reported that applied the framework proposed in the preceding chapters to the recognition of morphologically complex spoken words such as *builder* (*build+er*).

The second part takes up a central question in language learning: Can general regularities be acquired before specific ones, or do learners only posit a generalization once they have learned several specific instances of it (conservative generalization)? In nonprobabilistic models, conservative

generalization can avert commitment to overly broad generalizations that are difficult to retreat from. Three artificial language experiments on the learning of phonotactics (regularities in how sounds combine to form words) found that learners were not conservative: they generalized from a single item and were able to learn general rules without first learning their specific instances. A Bayesian model of phonotactic learning is proposed to account for these results. It considers both specific and general hypotheses simultaneously, but incorporates a parsimony bias: when less is known about the language, a single general pattern is acquired; with additional data, the bias is overcome and multiple specific patterns are learned. The model illustrates that conservative generalization is unnecessary in a probabilistic learner that maintains uncertainty about the rules of the language.

Contents

List of Figures

List of Tables

1 Introduction

Language speakers are faced with a formidable task. Most natural language rules have exceptions. The linguistic input is noisy: our interlocutors often pronounce words in a nonstandard or unfamiliar way. Probabilistic processing – the ability to entertain multiple possible hypotheses and ascribe a different degree of likelihood to each one – provides a way for learners and comprehenders to cope with the noisy input that they receive and leverage its statistical patterns to their advantage.

This dissertation investigates some of the consequences of evaluating multiple probabilistic hypotheses concurrently. It is divided into two parts. Part 1 (Chapters 2 through 5) addresses the impact of uncertainty over hypotheses in online comprehension. The linguistic input that we receive can have one or more interpretations. If there are very few possible analyses, or if one of the analyses is much more likely than the others, there is little uncertainty about the correct analysis of the input. Conversely, if a large number of potential analyses are equally likely, uncertainty is high. How does the degree of uncertainty affect comprehension difficulty, as reflected by reading times and neural activity?

Part 2 (Chapters 6 and 7) takes up generalization in language learning. Some models of generalization assume that learners are conservative: they learn generalizations that are as narrow as possible while still being compatible with the input. The experiments I present in this part show that this is empirically false – learners can learn broad patterns before narrow ones. I then present a probabilistic model that evaluates both narrow and broad generalizations concurrently and derives the initial preference for broad generalizations from a parsimony bias.

1.1 Uncertainty in comprehension

During language comprehension, multiple representations can be compatible with the input at any given time. This is most clearly the case for sentences that have several possible interpretations even after all of their words have been identified (Chomsky, 1965):

(1) Flying planes can be dangerous.

Even when a sentence has only one eventual analysis, its first few words are often temporarily compatible with multiple analyses. Consider, for example, the following two sentences (Marcus, 1980):

(2) a. Have the students take the exam today.

b. Have the students taken the exam today?

Each of the sentences has a single ultimate interpretation. After the words *have the students* have been read, however, the interpretation of these words is ambiguous between "make sure that the students..." and "is it the case that the students have...".

If readers do not only analyze the syntactic structure of the words they have read so far, but also predict what the upcoming words might be (DeLong, Urbach, & Kutas, 2005; Fruchter, Linzen, Westerlund, & Marantz, 2015), the range of competing hypotheses becomes even larger. For instance, after reading a sentence that begins with the words *she forgot*, a reader could predict a parse in which the complement of the verb is a noun phrase (as in *she forgot his birthday*), a parse with a sentential complement (as in *she forgot that his birthday was on Tuesday*), and so on.

Part 1 of the dissertation investigates how language processing is affected by the fact that multiple representations are considered at once. The assumption will be that readers or listeners associate every representation or interpretation with a probability estimate derived from the frequency of similar structures in the language (Jurafsky, 1996). For example, if a particular verb has appeared with a noun phrase complement seven out of ten times in the past, readers may estimate

that the probability that the current instance of the verb will be followed by a noun phrase is 0.7 (Trueswell, Tanenhaus, & Kello, 1993). Following earlier work, the main quantity of interest will be not the *number* of relevant hypotheses, but the *uncertainty* over the correct representation; the uncertainty may be higher in case there are two equally probable representations than in case there are ten possible representations, of which one is much more likely than the rest.

Chapter 2 reviews the framework of information theory that will be used to investigate this question, and discusses relevant previous work. Chapter 3 then investigates the role of uncertainty over syntactic parses in sentence processing. It reports on a self-paced reading study that found that words that reduce the reader's uncertainty about the syntactic structure of the sentence, as estimated from a corpus-based probabilistic grammar, can cause a slowdown in reading.

If all predicted syntactic continuations of a verb (its various complement types) are activated when the verb is read in the context of a sentence, can that activation be detected when the verb is read in isolation? Chapter 4 investigates this question using magnetoencephalography (MEG), a noninvasive technique that measures the magnetic fields generated by the electric activity in the brain. Higher uncertainty over the syntactic complements of the verbs was found to correlate with weaker neural activity in the left anterior temporal lobe, a brain region that has been associated with sentence-level processes.

Multiple concurrently activated representations are implicated in spoken word recognition as well. The sounds of a word are heard sequentially; listener incrementally activate the words that are compatible with the sounds they have heard to far, before they have encountered all of the sounds of the word (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987). Words are often made up of smaller units, called morphemes; *talking*, for example, can be seen as made up of *talk* and *-ing*. When the morpheme *talk* is identified, different potential upcoming morphemes can be predicted (e.g., *-ing*, *-s*, and *-ed*; see Ettinger, Linzen, & Marantz, 2014). Chapter 5 reports an MEG study that applies the framework of uncertainty over simultaneously activated representations, at both the segmental and morphological levels, to predict neural activity in auditory and language regions.

1.2 Rapid generalization in phonotactics

One of the defining properties of our knowledge of language is our ability to generalize beyond the specific items (sounds, words, sentences) that we have encountered in our linguistic input. This part of the dissertation focuses on generalization in phonotactics, the knowledge of which sequences of sounds are likely to be words of the language; for example, the knowledge that English words can start with an [s], or that *sliff* is a better potential word of English than *mpiff*.

How are these generalizations formed? One tradition in linguistics argues that learners form the smallest generalization that is consistent with their input; this is often referred to as the Subset Principle (Dell, 1981; Berwick, 1985; R. Clark & Roberts, 1993; M. Hale & Reiss, 2003). This tradition has led to the minimal generalization principle in phonotactics (Albright, 2009; Adriaans & Kager, 2010; see also Albright & Hayes, 2003). If a learner of a simple language has only encountered words that end in a [b], he or she will assume that all words must end in a [b]; once the learner has noticed that words can also end in [d], it will conclude that words can end in voiced stops in general (the minimal generalization that encompasses both [b] and [d]), but not in any other consonants.

Chapter 6 presents a series of artificial language learning experiments that evaluate the claim that humans are conservative learners. The experiments found that humans were able to learn broader generalizations without learning their narrower instances first; for example, they showed evidence of learning that words tended to have duplicated consonants (e.g., *bubu*) before showing evidence of recognizing any of the particular consonants that could be duplicated (e.g., distinguishing the consonant pair *bb*, which was in the input, from *dd*, which wasn't). They also generalized based a single item: having learned that words can begin with one particular voiceless stop (e.g., [k]), they generalized that knowledge to a different voiceless stop (e.g., [p]).

Conservative generalization strategies have their origin in nonprobabilistic learning frameworks. In those frameworks the scope of generalizations can only increase; as the learner cannot retreat from excessively broad generalizations, it is crucial to avoid generalizing beyond the data. Chapter 7 shows how this assumption is unnecessary in probabilistic learning. It presents

a Bayesian model that correctly captures the pattern of rapid, abstract generalization that human learners showed in the experiments. The learner entertains hypotheses that make reference to broad classes of sounds even after minimal exposure to the language. It prefers to learn compact or parsimonious explanations of the input corpus; this bias favors a single generalization over multiple narrower ones. The model does not become trapped in overgeneralizations: since it is probabilistic, it can entertain both narrow and broad hypotheses at the same time. As more linguistic input is received, probability shifts from the hypothesis assuming a single general pattern to the hypothesis assuming multiple more specific patterns; this allows the learner to retreat from overgeneralizations in a straightforward way, without requiring generalization to be conservative.

2 Entropy and information in human language processing

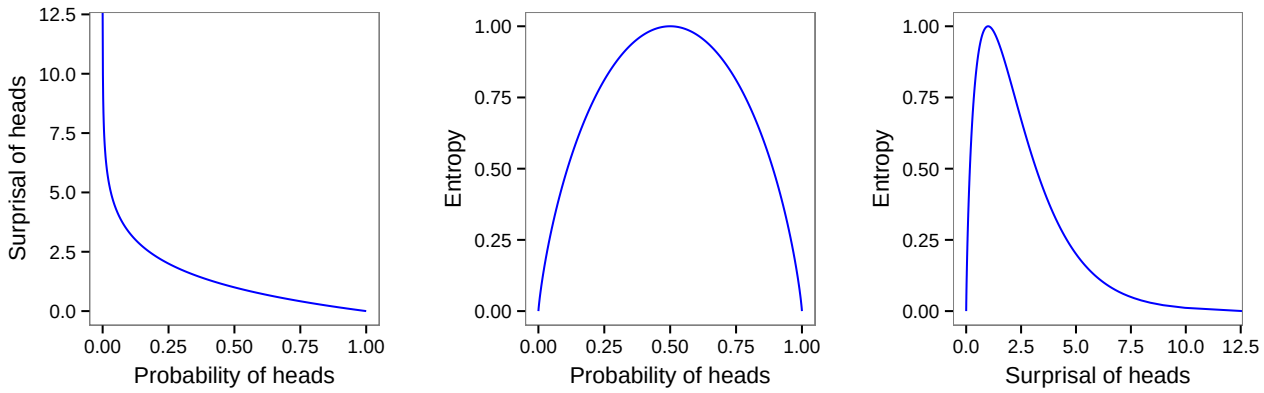
Entropy and related information theoretic measures are increasingly used to predict human behavior and neural activity, across different domains of psycholinguistics and with a variety of goals and theoretical motivations. This review aims to provide a unified introduction to this area, focusing on language comprehension. For reviews of other aspects of language use that have also been argued to be sensitive to information theoretic measures, such as speech production and typology, see Blevins (2013) and Jaeger and Tily (2011). For other introductions to basic concepts in information theory, see MacKay (2003), Manning and Schütze (1999) and Jurafsky and Martin (2008).

2.1 Basic concepts in information theory

In what follows, we use x to refer to a linguistic unit (word, suffix, syntactic structure and so on, depending on the context), and A to refer to the inventory of all possible units in the given context.

2.1.1 Surprisal

Suppose that the probability distribution P encodes our beliefs about the probabilities of an outcome (e.g., a probability distribution over words). The surprisal of each specific outcome x is



(a) Surprisal vs. probability (b) Entropy vs. probability (c) Entropy vs. surprisal

Figure 2.1: Relationship between surprisal, entropy and probability in a distribution with two outcomes (a biased coin).

defined as[1]

$$I(x) = \log_2 \frac{1}{P(x)} = -\log_2 P(x) \tag{2.3}$$

Surprisal is also commonly referred to as the Shannon information content (Shannon, 1948). The lower the probability of an outcome, the more surprising or informative it is considered to be. Since $P(x)$ is always 1 or less, and the logarithm of a number between 0 and 1 is 0 or less, $I(x)$ is always non-negative. $I(x) = 0$ only if $P(x) = 1$: if we know with certainty what the word is going to be, we do not gain any information from observing it. Due to the logarithmic relationship between surprisal and probability, the difference in surprisal between two outcomes with probabilities 0.01 and 0.05 is identical to the difference in surprisal between two outcomes with probabilities 0.1 and 0.5 (Figure 2.1a). Surprisal is often measured in bits.

[1]The following two properties of logarithms will be useful in the remainder of this review:

$$\log_2 \frac{a}{b} = \log_2 a - \log_2 b \tag{2.1}$$

$$\log_2 1 = 0 \tag{2.2}$$

2.1.2 Entropy

The entropy of a probability distribution P is the expected surprisal of its outcome: how surprising on average do we expect the outcome to be? Formally, it is defined as the weighted average of the surprisal of all outcomes, where the weights are given by the probabilities of the outcomes:

$$H = \sum_{x \in A} P(x) I(x) = -\sum_{x \in A} P(x) \log_2 P(x) \tag{2.4}$$

By convention, $0 \log_2 0 = 0$: outcomes with zero probability do not contribute to entropy. Like surprisal, entropy is measured in bits. Since surprisal is always non-negative, and entropy is expected surprisal, it follows that entropy is also non-negative.

Entropy can be seen as a measure of uncertainty about the outcome. If only one outcome has any probability of occurring—that is, there is a single outcome x_0 for which $P(x_0) = 1$—the entropy is zero: there is no uncertainty about the outcome. If the number of outcomes $n = |A|$ is held fixed, uncertainty is highest when all of the outcomes are equally likely:

$$H(P) = -\sum_{x \in A} \frac{1}{n} \log_2 \frac{1}{n} = -\frac{n \log_2 \frac{1}{n}}{n} = \log_2 n \tag{2.5}$$

Equation 2.5 can also be interpreted as showing that if the distribution across outcomes is uniform, entropy increases with the number of outcomes. For two equiprobable outcomes the entropy will be one bit; for four equiprobable outcomes, it will be two bits.[2] If the number of outcomes is kept constant, the entropy of the distribution decreases the more uneven (peaked) the distribution is. If there are two outcomes x_1 and x_2, our uncertainty about the outcome is higher when they are equally likely than when one of which is much more likely than the other. For example, if $P(x_1) = 0.9$ and $P(x_2) = 0.1$, then $H(P) = 0.47$, less than half of the entropy of the case of two equally probable outcomes (see Figures 2.1b and 2.2).

[2]This observation may provide an intuition for why entropy is measured in bits: a variable that has four outcomes can be efficiently coded using two **b**inary dig**its** (00, 01, 10, and 11).

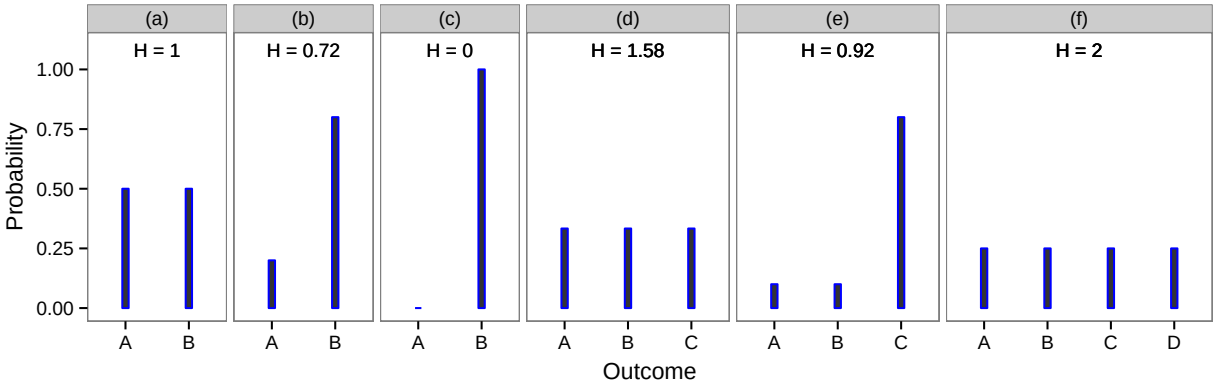


Figure 2.2: Entropy of some sample distributions: (a) Two equiprobable outcomes; (b) Two outcomes, one more probable than the other; (c) Two outcomes, only one of which is possible; (d) Three equiprobable outcomes; (e) Skewed distribution over three outcomes: H is lower than in the case of two equiprobable outcomes; (f) Four equiprobable outcomes (when outcomes are equiprobable, H is logarithmic in the number of outcomes).

2.1.3 Relative entropy (KL divergence)

The relative entropy (Kullback-Leibler divergence) of a probability distribution P with respect to another distribution Q is a measure of how well Q approximates P. It is defined as the expectation of the logarithm of the likelihood ratio between the two distributions, when the expectation is taken with respect to P:[3]

$$D(P \parallel Q) = \sum_{x \in A} P(x) \log_2 \frac{P(x)}{Q(x)}$$

By convention, $0 \log_2 \frac{0}{0} = 0$ (if both distributions assign a zero probability to an outcome, that outcome does not contribute to the relative entropy) and $p \log \frac{p}{0} = \infty$ (relative entropy is infinite if there is an outcome assigned zero probability by Q but not by P).

While for each specific x the log-likelihood ratio $\log_2 \frac{P(x)}{Q(x)}$ can be either positive or negative, overall KL divergence is guaranteed to be nonnegative, with equality holding only when $P = Q$ (Gibbs' inequality; see MacKay, 2003, p. 34). KL divergence is often referred to as a distance

[3]Note that in contrast to the definition of entropy there is no minus sign in front of the summation.

w	$P(w)$	$P(w\|[\text{k}])$	$P(w\|[\text{kæ}])$
dog	$2/10$	0	0
cat	$5/10$	$5/8$	1
koala	$3/10$	$3/8$	0
	$H = 1.48$	$H = 0.95$	$H = 0$

Table 2.1: Language A. $P(w)$ is the prior probability of w (typically, the relative frequency of w in the language); $P(w|[\text{k}])$ is the probability of w given that the first phoneme is [k].

measure, even though it does not have some of the standard properties of a distance measure. Most importantly, it is typically not symmetrical: in most cases $D(P \parallel Q) \neq D(Q \parallel P)$.

2.2 Information in sequential processing

In typical language comprehension scenarios, the comprehender attempts to infer the identity of a larger linguistic unit, such as a word or a sentence, from sequentially perceived elements. In the case of spoken word recognition, for example, each phoneme might change the listener's beliefs about the identity of the word being perceived. It is often of interest to measure how much information about the larger unit is conveyed by each of the elements. The rest of this discussion will illustrate the application of information theory to sequential comprehension, using the example of sequential recognition of phonemes in a spoken word. The discussion for word-by-word sentence comprehension is conceptually analogous; however, simple examples in that domain are harder to construct because the processes that are assumed to generate words in sentences (e.g., context-free grammars) are more complex than those that generate phonemes (a finite lexicon). Section 2.4.1 discusses some of the challenges posed by surprisal and entropy calculations in infinite languages.

2.2.1 Sequential processing: a simple example

We will illustrate the process of phoneme-by-phoneme spoken word comprehension using Language A (Table 2.1). The language includes three words, *dog*, *cat* and *koala*. The prior probabilities of those words are given by $P(w)$: $2/10$, $5/10$ and $3/10$, respectively; in practice these probabilities

are often estimated from the relative frequency of the words. When the listener discovers that the identity of the first phoneme Φ_1 is ϕ_1, his or her beliefs are expressed by the distribution $P(w|\phi_1)$. For example, if the first phoneme is [k], the probability that the word is *dog* drops to zero, and the probability that the word is *cat* or *koala* increases. The second phoneme causes a further update in the listener's beliefs about the identity of the word—in this case, identifying it unambiguously.

2.2.2 Two measures of sequential information gain

There are two standard ways to quantify the amount of information that a particular value of Φ_1 conveys about W. One is based on the KL divergence of the probability distribution before and after conditioning on the outcome Φ_1, that is

$$I^{\text{KL}}(\phi_1) = D((P|\Phi_1 = \phi_1) \,||\, P)$$

A phoneme that causes a larger update in the listener's beliefs is seen as more informative about the identity of the word. Another way to quantify the information conveyed by Φ_1 about the word is using *entropy reduction* (K. Wilson & Carroll, 1954)—the difference between the uncertainty about the identity of the word before and after ϕ_1 was perceived:

$$I^{\text{ER}}(\phi_1) = H(P) - H(P|\Phi_1 = \phi_1)$$

Under the latter definition, phonemes that reduce uncertainty to a greater extent are seen as more informative. The two measures of information gain are in general not equal. Consider the example of the sequential perception of *cat* in the language described above. The sequential en-

w	$P(w)$	$P(w\|[\text{k}])$	$P(w\|[\text{kæ}])$
dog	0.772	0	0
cat	0.114	0.5	1
koala	0.114	0.5	0
	$H = 1$	$H = 1$	$H = 0$

(a) Language B

w	$P(w)$	$P(w\|[\text{k}])$	$P(w\|[\text{kæ}])$
dog	0.9	0	0
cat	0.05	0.5	1
koala	0.05	0.5	0
	$H = 0.57$	$H = 1$	$H = 0$

(b) Language C

Table 2.2: Entropy can remain constant (Language B) or increase (Language C), even when there is a change in beliefs.

tropy values are as follows:

$$\begin{aligned} H(P) &= -(2/10 \log_2 2/10 + 5/10 \log_2 5/10 + 3/10 \log_2 3/10) = 1.48 \\ H(P|[\text{k}]) &= -(5/8 \log_2 5/8 + 3/8 \log_2 3/8) = 0.95 \\ H(P|[\text{kæ}]) &= -1 \times \log_2 1 = 0 \end{aligned} \tag{2.6}$$

The entropy reduction caused by [k] is therefore $1.48 - 0.95 = 0.53$; the entropy reduction caused by [æ] is $0.95 - 0 = 0.95$.

The relative entropy values, on the other hand, are[4]

$$I^{\text{KL}}([\text{k}]) = 5/8 \log_2 \frac{5/8}{5/10} + 3/8 \log_2 \frac{3/8}{3/10} = 0.32 \tag{2.7}$$

$$I^{\text{KL}}([\text{æ}]|[\text{k}]) = 1 \times \log_2 \frac{1}{5/8} = 0.56 \tag{2.8}$$

Under the entropy reduction definition of informativity in sequential processing, the information conveyed by the second phoneme is greater than the information conveyed by the first one; the opposite holds for the definition based on KL divergence.

[4]Outcomes that have zero probability under the "new" distribution do not contribute to KL divergence and can be dropped.

I^{KL} is always nonnegative; it equals zero only if the listener's beliefs didn't change at all upon hearing the last phoneme. This is not the case for I^{ER}. Consider, for example, the sequential processing of *cat* in Language B, which has the same words as Language A but a different probability distribution over those words. In Language B, *cat* and *koala* are equally likely, but *dog* is more likely than either of those words. Although there are three possible words, the distribution is skewed in favor of *dog*, and the entropy is only one bit. If the first phoneme is [k], the word *dog* is no longer a possible candidate. The two words that remain compatible with the input are equally likely. The uncertainty about the outcome is still one bit, the maximal uncertainty possible in a distribution with two outcomes. The entropy reduction caused by [k] is therefore zero: despite the change in beliefs, the entropy reduction metric of informativity does not ascribe any informativity to [k] in this context.

Entropy reduction can also be negative. Language C is similar to Language B, but has an even stronger skew towards *dog*. Listeners can be fairly confident that the word will be *dog* before they even begin to hear the word; consequently, uncertainty is low ($H = 0.57$). When the first phoneme turns out to be [k] and *dog* becomes irrelevant, there is again maximal uncertainty over the remaining two options ($H = 1$). The entropy reduction caused by [k] is therefore -0.43.

To avoid negative entropy reduction values, a phoneme for which $I^{\text{ER}} < 0$ can be defined as not providing any information about the word (J. Hale, 2006; Yun, Chen, Hunter, & Hale, 2015). In other words, entropy reduction can be replaced by the nonnegative measure $\hat{I}^{\text{ER}}$, defined as

$$\hat{I}^{\text{ER}} = \max\{I^{\text{ER}}, 0\}$$

Even with this adjustment, however, it is arguably undesirable for a measure of information gain to be equal to zero when the phoneme has changed the listener's beliefs about the word (Blachman, 1968).

w	$P(w)$	$P(w\|[\text{k}])$	$P(w\|[\text{kæ}])$	$P(w\|[\text{ko}])$
dog	$2/10$	0	0	0
cat	$5/10$	$5/8$	1	0
koala	$3/10$	$3/8$	0	1
	$H = 1.48$	$H = 0.95$	$H = 0$	$H = 0$
			$P([\text{kæ}]\|[\text{k}]) = 5/8$	$P([\text{kæ}]\|[\text{k}]) = 3/8$

Table 2.3: The entropy reduction caused by an outcome is not necessarily related to its conditional probability (illustrated for Language A).

2.2.3 Relationship between sequential information gain and surprisal

Reduction in uncertainty caused by a phoneme is not necessarily related to the conditional probability of the phoneme. Going back to Language A, if the two first phonemes of the word are $\Phi_1 = [\text{k}]$ and $\Phi_2 = [\text{o}]$, the identity of the word has to be *koala*; likewise, if $\Phi_1 = [\text{k}]$ and $\Phi_2 = [\text{æ}]$, the word is necessarily *cat*. In neither case is there any uncertainty left about the identity of the word after the two phonemes are perceived (Table 2.3). Since the entropy before and after the second vowel is the same in both cases, I^{ER} on the second phoneme will be identical, even though [o] is more surprising than [æ].

Under the KL divergence definition of information gain, on the other hand, the less probable phoneme [o] is considered to be more informative:

$$I^{\text{KL}}(\text{o}|\text{k}) = 1 \times \log_2 \frac{1}{3/8} = 1.41 \tag{2.9}$$

$$I^{\text{KL}}(\text{æ}|\text{k}) = 1 \times \log_2 \frac{1}{5/8} = 0.56 \tag{2.10}$$

In fact, I^{KL} at the i-th phoneme Φ_i is always equal to that phoneme's surprisal given the prefix $\Phi_1, \ldots, \Phi_{i-1}$. This equivalence holds under the very general assumption that the process that generates the possible full forms remains constant throughout the perception of the form (Levy, 2008a). Section 2.2.4 includes a detailed discussion of this equivalence.

2.2.4 More on the relationship between relative entropy and surprisal

This section elaborates on the equivalence between relative entropy and surprisal in sequential processing mentioned in Section 2.2.3; readers who are not interested in this equivalence can safely skip this section.

In Language A, the first phoneme is [d] with probability $2/10$ (only *dog* starts with a [d]) and [k] with probability $8/10$ (the sum of the probabilities of *cat* and *koala*). Here the surprisal of [k] is $-\log_2 8/10 = 0.32$, again equal to the result of Equation 2.7, repeated here:

$$I^{\text{KL}}([\text{k}]) = 5/8 \log_2 \frac{5/8}{5/10} + 3/8 \log_2 \frac{3/8}{3/10} = 0.32 \tag{2.11}$$

Next, given that the first phoneme was [k], the second phoneme is [o] with probability $3/8$ and [a] with probability $5/8$. The surprisal of [o] in this context is $-\log_2 3/8 = 1.41$, the same quantity as $I^{\text{KL}}(\text{o}|\text{k})$ in Equation 2.9. Note that both of the likelihood ratios in Equation 2.11 are equal to the inverse of $P(\Phi_1 = [\text{k}])$:

$$\frac{5/8}{5/10} = \frac{3/8}{3/10} = \frac{1}{8/10} = \frac{1}{P(\Phi_1 = [\text{k}])} \tag{2.12}$$

We can therefore collect terms and simplify, illustrating the equivalence between KL divergence and surprisal:

$$I^{\text{KL}}([\text{k}]) = (5/8 + 3/8) \log_2 \frac{1}{P(\Phi_1 = k)} = -\log_2 P(\Phi_1 = [\text{k}]) \tag{2.13}$$

When the identity of the next phoneme is revealed, some of the words in the language are typically ruled out; the listener "zooms in" on a subset of the lexicon. Yet the relative probabilities of the options that are still relevant remain the same: *cat* is $8/5$ times more likely than *koala,* regardless of what the first phoneme was.[5] When some of the total probability mass is lost to the

[5]This is true even when the first phoneme was not [k], and consequently both of the words have zero probability.

options ruled out by the incoming phoneme, the probabilities of the surviving options need to be renormalized so that they all sum to 1 again. The renormalization factor is equal to the sum of the probabilities that the surviving options had before the incoming phoneme was encountered; the sum is equal to the probability of the incoming phoneme. As an example, suppose that the first phoneme in the input is [k]. We have

$$I^{\text{KL}}([\text{k}]) = \sum_{w \in A} P(W = w|\Phi_1 = [\text{k}]) \log_2 \frac{P(W = w|\Phi_1 = [\text{k}])}{P(W = w)} = \sum_{w \in A} P(W = w|\Phi_1 = [\text{k}]) \log_2 \frac{\frac{P(W=w, \Phi_1=[\text{k}])}{P(\Phi_1=[\text{k}])}}{P(W = w)} \tag{2.14}$$

Let $\phi_1(w)$ be the first phoneme of w for every word w in the lexicon; then

$$P(W = w, \Phi_1 = [\text{k}]) = \begin{cases} P(W = w) & \text{if } \phi_1(w) = [\text{k}] \\ 0 & \text{otherwise} \end{cases}$$

Plugging this quantity back into Equation 2.14, we can now discard all of the words that do not start with [k], as their contribution to I^{KL} is 0, and limit ourselves to words that start with [k] (denoted here $w|\phi_1(w) = [\text{k}]$):

$$= \sum_{w|\phi_1(w)=[\text{k}]} P(W = w|\Phi_1 = [\text{k}]) \log_2 \frac{\frac{P(W=w)}{P(\Phi_1=[\text{k}])}}{P(W = w)} \tag{2.15}$$

The term $P(W = w)$ in the numerator and denominator cancels out; the remaining term $\log_2 \frac{1}{P(\Phi_1=[\text{k}])}$ does not depend on w and can therefore be taken outside the summation:

$$= \sum_{w|\phi_1(w)=[\text{k}]} P(W = w|\Phi_1 = [\text{k}]) \log_2 {}^1/_{P(\Phi_1=[\text{k}])} = \log_2 {}^1/_{P(\Phi_1=[\text{k}])} \sum_{w|\phi_1(w)=[\text{k}]} P(W = w|\Phi_1 = [\text{k}]) \tag{2.16}$$

Note that the summation is now over the probabilities given [k] of all words that start with [k]; this quantity is equal to 1:

$$= \log_2 {}^1\!/_{P(\Phi_1=[\text{k}])} \times 1 = I([k]) \tag{2.17}$$

This shows the equivalence between I^{KL} and the surprisal of the incoming phoneme.

2.2.5 Sensitivity to representational assumptions

The equivalence between relative entropy and the surprisal of the incoming surface element limits the sensitivity of relative entropy to assumptions about abstract structure. This is never an issue with the simple lexicon models we used to illustrate the concepts in Section 2.2 for the case of phoneme-by-phoneme spoken word recognition. In those models, a particular surface string is generated by exactly one underlying representation. For example, the phoneme sequence [d], [ɑ], [g] can only be generated by *dog*. When language models with more abstract representations are used, however, the same surface string may be generated by several different abstract representations. The range of ambiguous abstract representations may affect entropy-based information, but will not affect relative entropy.

For example, suppose that our goal is to estimate the informativity of the fourth phoneme [ɪ] in the surface form *kicking* ([kɪkɪŋ]). We assume that only three continuations of *kick* are possible, and that all three are equally probable (see Table 2.4). Now consider two language models. In M_{surface}, the lexicon consists of whole word forms, and the two representations for *kicks* are collapsed. Under this model, the probability distribution over full representations after *kick* has been processed is ${}^1\!/_3$ for *kick* and ${}^2\!/_3$ for *kicks*. M_{morph}, by contrast, represents the full morphological analysis of the words. Consequently, while the probability of *kicking* is still ${}^1\!/_3$, it competes with two other representations, *kicks (verb)* and *kicks (noun)*. Each of the two forms (which correspond to the same surface form) is expected with probability ${}^1\!/_3$. The entropy of the distribution over representations is higher in M_{morph} than in M_{surface}, and is therefore reduced to a greater extent when

Form	Morphological analysis	Frequency	P_{morph}	P_{surface}
kicking	*kick* (verb) + present participle	100	1/3	1/3
kicks	*kick* (verb) + third person present	100	1/3	2/3
	kick (noun) + plural	100	1/3	
			$H = 1.58$	$H = 0.92$

Table 2.4: A simplified morphological paradigm for *kick* with two syncretic forms (homographic inflected forms that have different morphological analyses). P_{morph} is a probability of the inflected form given that the stem is *kick* under a model that is sensitive to morphological structure, and P_{surface} is the same probability under a model that is only sensitive to the word's surface form.

the phoneme [ɪ] is revealed and eliminates the listener's uncertainty about the identity of the word. Yet the relative entropy is the same in both cases—the representations that are no longer compatible with the string now have zero probability, and therefore do not contribute to relative entropy. As long as the total probability mass assigned to the eliminated representations is identical, so too will the relative entropy be.

The sensitivity of entropy to representational assumptions makes it a potentially useful tool for distinguishing two theories of abstract representations (J. Hale, 2006). Given a linking function between entropy and behavioral measures, entropy estimates that predict the dependent measures well can serve as evidence for a particular language model. On the other hand, if entropy is used as a predictor in a context where representational assumptions are not of theoretical interest, it may be prudent to verify that the results are robust to those assumptions.

2.2.6 Sources of uncertainty

In this review we assume that there is no uncertainty about the input received so far: in the case of spoken word recognition, the only sort of uncertainty is about the identity of the word out of all words that are compatible with the first few phonemes. Whenever listeners perceive a phoneme, they can be confident that they have perceived it veridically and no longer need to maintain hypotheses about words in the lexicon that do not begin with the sequence of phonemes they have perceived. This assumption is particularly associated with the cohort model of spoken word pro-

cessing (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), but is likely to be false. For example, listeners can understand the word *cigarette* when it is mispronounced by a drunk person as *shigarette*; the cohort model predicts that *cigarette* should drop out of the cohort immediately when the first phoneme of *shigarette* is revealed to be incompatible with *cigarette*. A natural interpretation of this phenomenon is that in the process of spoken word recognition listeners always assign a nonzero (though potentially very low) probability to all words in the lexicon (Norris, 1982; Dahan, Magnuson, & Tanenhaus, 2001; for a review, see Dahan, 2010). A similar argument has been made for sentence processing (Levy, 2008b; Bicknell & Levy, 2010). Combining estimates of the informativity of an incoming phoneme with the uncertainty about past phonemes is an area of active research (Levy, 2008b, 2011). As information metrics that incorporate past uncertainty are developed, this simplification should be revisited.

2.3 Linking hypotheses

This section surveys proposals about the possible effects of surprisal and entropy on cognitive dependent measures. To simplify the exposition, we will sometimes abstract away from the particular dependent measures (reading times, various neural signals) and use the general term "processing load". In practice, these variables may affect different dependent measures in different ways; we discuss some cases in which that is the case. Moreover, the hypotheses are not mutually exclusive: different processes may be sensitive to different measures, and the same process may be sensitive to multiple measures at the same time (Frank, 2013).

2.3.1 Surprisal and probability

The proposal that processing load on a unit in sequential processing is proportional to its surprisal was advanced in J. Hale (2001), following earlier work by Attneave (1959). Since most of the discussion in the literature focuses on word surprisal in sentence processing, we will use that case as our example.

The statement that processing load is affected by the information conveyed by a word is an abstract ("computational-level") statement. What concrete cognitive model might give rise to a correlation between the surprisal of a word and processing load at the word (see also Levy, 2013)? J. Hale (2001) proposes that the processing cost associated with surprising words arises from the disconfirmation of all of the predicted sentences (parses) that turned out to be incompatible with w_n. Indeed, if the set of sentences disconfirmed by w_i is $\{T_i^n\}$, then the surprisal of w_n is equal to

$$\log_2 \left(\frac{1}{1 - \sum P(T_i^n)} \right)$$

The cost of disconfirmation is assumed to be determined by the total probability of the disconfirmed sentences. It is unaffected by their distribution: disconfirming a single sentence whose probability is 0.9 is just as costly as disconfirming nine sentences whose probability is 0.1 each. There is no independent empirical evidence that this is in fact the case, and one might imagine measures of disconfirmation effort that do not have this property. In some architectures, for example, the cost of disconfirming a sentence might be fixed; in that case, the disconfirmation cost incurred by w_i would simply be the number of disconfirmed hypotheses, $|\{T_i^n\}|$.

An alternative interpretation of the effect of surprisal is based on the equivalence of the surprisal of a word on the one hand and the relative entropy between the distribution over sentences before and after the word has been processed on the other hand (see Section 2.2.3). This interpretation of surprisal effect rests on the assumption that larger changes in beliefs are more costly. While this is a plausible assumption, it is unclear whether there are empirical findings that directly support it, and there are many other possible cognitive architecture; for example, one might imagine a system in which the cost of a probability update would depend on the number of cells in the distribution whose probability needs to be updated (this would be the case in a straightforward implementation on a standard computer).

Finally, in the rational framework proposed by Norris (2006) and Smith and Levy (2013), the higher processing cost associated with higher surprisal is due to probability-sensitive preactivation:

if an upcoming word has high conditional probability given what is known so far about the sentence, the reader will allocate more resources to processing it, in advance of reading it (Fruchter et al., 2015). A high-surprisal incoming word will not have been preactivated to the same extent as a low-surprisal word, and will therefore not enjoy the speedup conferred by preactivation.

Although the preactivation account offers a straightforward reason for expecting a predictability effect (i.e., an association between conditional probability and processing cost), it does not derive the logarithmic shape of this effect. Indeed, many studies have assumed a linear rather than logarithmic link between conditional probability and dependent measures (e.g., cloze probability in DeLong et al., 2005 or "prediction error" in Gagnepain, Henson, & Davis, 2012). These two functions can make very different predictions, especially for low conditional probabilities. This is illustrated in Language E (Table 2.5; see also Figure 2.1a). In this language, the probability difference between [a] and [e] is equal to the probability difference between [e] and [i]:

$$P([\text{e}]) - P([\text{i}]) = P([\text{a}]) - P([\text{i}]) = 0.099$$

Yet the difference in surprisal between [e] and [i] is much higher than the difference in surprisal between [a] and [e]:

$$\begin{aligned} I([\text{e}]) - I([\text{i}]) &= 1 \\ I([\text{a}]) - I([\text{e}]) &= 9.97 \end{aligned}$$

Recent empirical studies suggest that the relationship between conditional probability and processing load is much closer to logarithmic than to linear, at least in the case of reading times, supporting the surprisal hypothesis (Smith & Levy, 2013).

Φ	$P(\Phi)$	$I(\Phi)$
a	0.0001	13.29
e	0.1	3.32
i	0.1999	2.32
u	0.7	0.51

Table 2.5: Language E: probability (P) vs. surprisal (I). Equal differences on the probability scale do not translate to equal differences on the surprisal scale.

2.3.2 Entropy reduction

The Entropy Reduction Hypothesis states that processing load on a word correlates with the degree to which the word reduced the listener's uncertainty about the identity of the sentence (J. Hale, 2003a, 2003b, 2006; Yun et al., 2015). No processing work is predicted when uncertainty increases. Whereas surprisal is fully determined by the total probability of alternatives ruled out by the incoming unit, entropy reduction depends on the distribution of the alternatives that were ruled out, as well as on the probabilities of the surviving options. This is illustrated in Language F (Table 2.6). The conditional probability of [a] is ¼ after either [t] or [b]; the surprisal of the second phoneme is therefore identical in *ta* and *ba*. At the same time, the remaining probability mass (¾) is distributed differently in the two cases: following [t], [ta] competes with three alternatives, each with probability ¼; following [b], [ba] only competes with a single alternative that has probability ¾. The entropy reduction caused by the second phoneme [a] is therefore larger after [t] (2 bits) than after [b] (0.81 bits).

Entropy reduction is often interpreted as disambiguation work; readers "are performing the maximum amount of disambiguation, an amount of work proportional to the information conveyed by the word" (J. Hale, 2003b, p. 119; see also J. Hale, 2006, p. 650). It is not immediately obvious why disambiguation difficulty should depend on the probabilities of the alternatives being ruled out, however. In Language F, it was intuitively clear why [a] was doing more disambiguation work when it ruled out three different words than when it rules out a single word. In Language G (Table 2.7), on the other hand, the phoneme [a] disconfirms exactly one option after either [b]

Word	$P(w)$	$P(w\|[\text{t}])$	$P(w\|[\text{ta}])$
ta	$1/8$	$1/4$	1
ti	$1/8$	$1/4$	0
te	$1/8$	$1/4$	0
tu	$1/8$	$1/4$	0
ba	$1/8$	0	0
bu	$3/8$	0	0
	$H = 2.4$	$H = 2$	$H = 0$

(a) Processing *ta*

Word	$P(w)$	$P(w\|[\text{b}])$	$P(w\|[\text{ba}])$
ta	$1/8$	0	0
ti	$1/8$	0	0
te	$1/8$	0	0
tu	$1/8$	0	0
ba	$1/8$	$1/4$	1
bu	$3/8$	$3/4$	0
	$H = 2.4$	$H = 0.81$	$H = 0$

(b) Processing *ba*

Table 2.6: Language F. The shaded cells indicate the conditional probability that the second phoneme will be a.

w	$P(w)$	$P(w\|[\text{t}])$	$P(w\|[\text{ta}])$
ta	$1/5$	$1/2$	1
tu	$1/5$	$1/2$	0
ba	$1/5$	0	0
bu	$2/5$	0	0
	$H = 1.92$	$H = 1$	$H = 0$

(a) Processing *ta*

w	$P(w)$	$P(w\|[\text{b}])$	$P(w\|[\text{ba}])$
ta	$1/5$	0	0
tu	$1/5$	0	0
ba	$1/5$	$1/3$	1
bu	$2/5$	$2/3$	0
	$H = 1.92$	$H = 0.92$	$H = 0$

(b) Processing *ba*

Table 2.7: Language G: Illustration of the probability-dependence of entropy reduction and its relationship to disambiguation work.

or [t]. Yet the entropy reduction hypothesis predicts that disambiguation in favor of *ta* would be harder than disambiguation in favor of *ba* because its competitor had higher probability. Echoing our discussion of the disconfirmation interpretation of surprisal, one could imagine a system in which disambiguation difficulty was only sensitive to the number of disconfirmed words.

As illustrated in Section 2.2.2, disconfirming some of the options can often lead to an *increase* in entropy (e.g., Language C in Table 2.2). The entropy reduction hypothesis does not predict any processing load in these cases, even though some options have been ruled out and intuitively some disambiguation work has been done.

Language H (Figure 2.3) illustrates the complex range of numerical predictions made by the entropy reduction hypothesis by systematically exploring a family of simple distributions over two-phoneme-long lexicons. The language includes the words *ba* and *ta*, as well as between one and three additional words that start with [t] (e.g., *ti*). We focus on the entropy reduction profile associated with the recognition of the word *ta*; we denote the entropy before [t] by h_1 and the entropy after [t] by h_2. Three points stand out in Figure 2.3:

1. When the word *ba*, which is ruled out by the first phoneme, has high probability ($p = 0.75$), it is often the case that $h_1 < h_2$. Since the entropy reduction caused by [t] is $h_1 - h_2$, the entropy reduction hypothesis does not attribute any disambiguation work to the first phoneme in these cases.

2. The second phoneme [a] uniquely identifies the word (entropy drops to 0); the entropy reduction it causes is equal to h_2. In general, the amount of disambiguation work prompted by the second phoneme is related to the number of alternatives ruled out by it—the curve is higher the more competitors *ta* had before the second phoneme [a] was revealed. At the same time, the elimination of a single alternative when q is 0.5 can result in higher entropy reduction than the elimination of three alternatives when q is 0.9.

3. The inverted U-shaped curves show that the relationship between the phoneme's surprisal q and the entropy reduction caused by it is not straightforward; in particular, when there is a single alternative [t]-initial word, more disambiguation work is predicted when q is 0.5 than when it is either 0.1 or 0.9.

In some studies entropy reduction is not capped at zero; Frank (2013), for example, considers negative entropy reduction values as well. The theoretical interpretation of such values is unclear. Empirically, whether or not entropy reduction can be negative is an important decision, since entropy often increases in practice (Frank, 2013; Linzen & Jaeger, 2014).

2.3.3 Competition

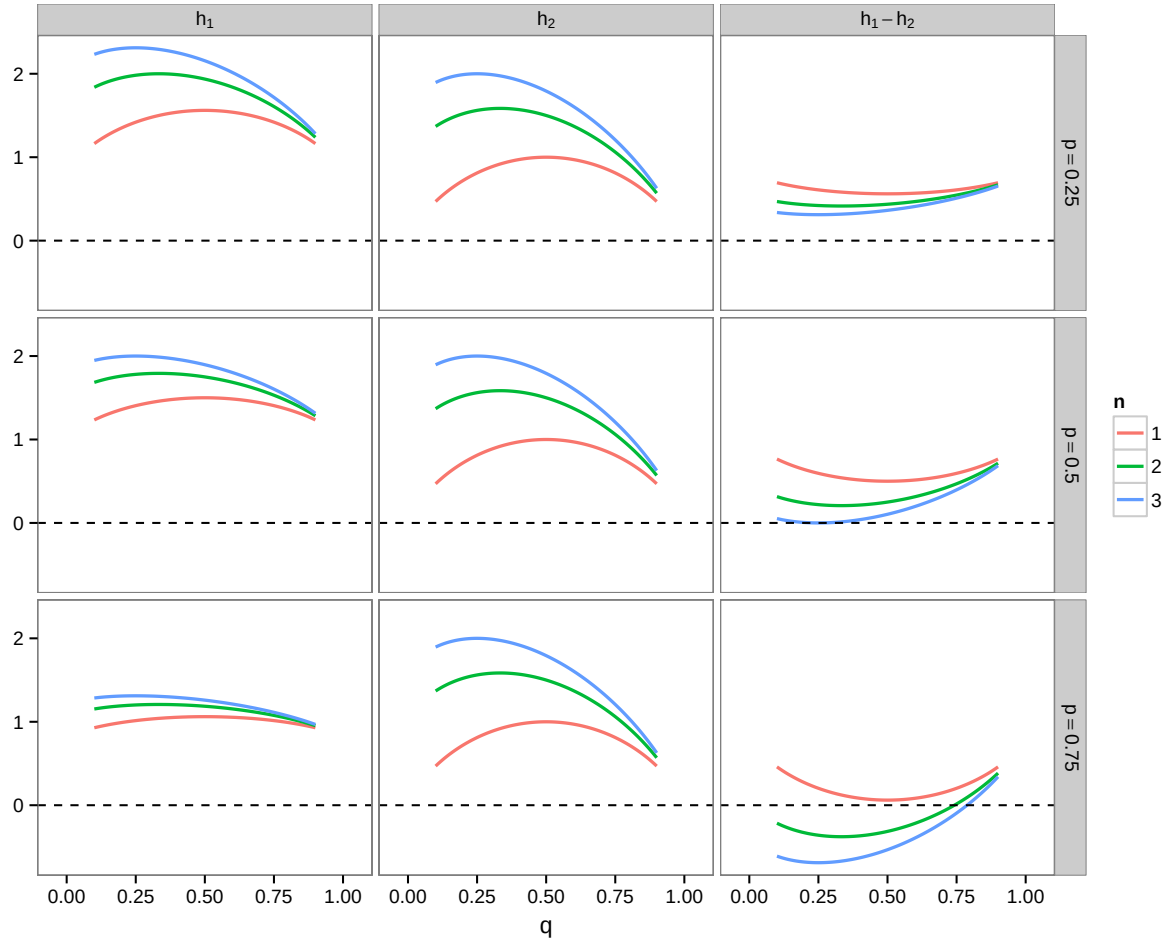
The discussion so far has assumed that all of the representations that are compatible with the input are simultaneously activated. If there is a linear relationship between the probability of a representation and the degree to which it is activated, total activation will be constant, since the total probability mass must always sum to one. In other words, if $p(w)$ denotes the probability of a word at a particular point and $a(w)$ its activation at that point, and if $a(w) = cp(w)$ (where c is a constant), then

$$\sum_w a(w) = \sum_w cp(w) = c\sum_w p(w) = c \tag{2.18}$$

One might imagine an architecture in which neural and behavioral measures were not affected by the distribution over currently active representations, but only by changes in that distribution

w	$P(w)$	$P(w\|[\text{t}])$	$P(w\|[\text{ta}])$
ba	p	0	0
ta	$q(1-p)$	q	1
ti_1	$(1-q)(1-p)/n$	$(1-q)/n$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
ti_n	$(1-q)(1-p)/n$	$(1-q)/n$	0
	$H = h_1$	$H = h_2$	$H = 0$

(a)



(b)

Figure 2.3: (a) Language H; (b) Entropy and entropy reduction profiles associated with the word *ta* in Language H, under various combinations of the parameters $p = P(ba)$, $q = P(ta|[\text{t}])$, and n (the number of [t]-initial words other than *ta*). The entropy reduction caused by [t] is $h_1 - h_2$. Since the entropy after [a] is always 0, the entropy reduction caused by [a] is equal to h_2.

as additional input was received (as proposed by the surprisal and entropy reduction hypotheses). There is evidence, however, that simultaneous activation of multiple representations leads to competition among those representations, affecting reaction times or neural activity. Intuitively, competition is stronger the more candidates there are and the more similar their probabilities are to each other; the entropy of the distribution over representations is therefore a natural candidate for quantifying the degree of competition among them (Gagnepain et al., 2012; Schmidtke, Kuperman, Gagné, & Spalding, 2015). In Language F, for example, when the first phoneme of the word is [t], four words are active with equal probability; when the first phoneme is [b], only two words are active. In this scenario, the competition hypothesis predicts increased competition after [t] has been processed than after [b].

The effect of competition is likely to depend on the nature of the competing representations as well as on the task that the listener is performing. In lexical decision experiments, for example, competition among the complex forms that are derived from the word being recognized can lead to faster recognition. In a lexical decision experiment the goal of the participant is to determine whether or not the input is a word; the participant is not required to identify which particular word the input makes up. In that scenario, a greater amount of activation of related representation that are consistent with the input may lead to faster recognition: although each representation has different semantic or morphological features, they all support the hypothesis that the input makes up a word (Baayen, Lieber, & Schreuder, 1997; Moscoso del Prado Martín, Kostić, & Baayen, 2004; see Section 2.5.2 below). On the other hand, when a reader is trying to understand a sentence, reading times may be longer if several different parses of the sentence are equally likely (McRae, Spivey-Knowlton, & Tanenhaus, 1998). Likewise, when a listener is attempting to identify the meaning of a spoken word, a word-initial phoneme prefix that is compatible with a large number of words may lead to slower recognition times than a prefix that uniquely identifies a single word.

Predicting the direction of competition effects is further complicated by the fact that competition among semantic interpretations sometimes leads to increased activation and a slowdown in reaction times across tasks (Duffy, Morris, & Rayner, 1988; Simon, Lewis, & Marantz, 2012).

This suggests that participants do sometimes create a semantic interpretation for the word even in a lexical decision experiment. Moreover, the direction of competition effects is sensitive to the similarity among the candidates competing for selection (for discussion and simulations, see Q. Chen & Mirman, 2012).

2.3.4 Commitment

While it is often assumed that readers and listeners activate all of the representations that are compatible with the input, it may occasionally be beneficial to only activate representations when there is relatively low uncertainty about the appropriate representation. This may be the case if there is a fixed processing cost associated with constructing a representation, if there is a limited number of representations that a particular cognitive process can keep track of, if processing is slowed down by competition, and so on. It might be the case that some cognitive processes, such as lexical access, can keep track of the full probability distribution, whereas other processes (e.g., a discourse level representations) have limited bandwidth and need to follow the commitment approach.

If there are thousands of words in the lexicon that have [t] as their first phoneme, for example, the listener may prefer to delay accessing the semantic representations of those words until more of the input is revealed and the number of candidates becomes more manageable. As such, the commitment hypothesis predicts that lower uncertainty will be associated with increased activation (cf. the Low Entropy Dependent Prediction hypothesis of Ettinger et al., 2014). Commitment to one or a few representations may result in revision cost if none of those representations turns out to be correct (Federmeier, 2007; Van Petten & Luka, 2012); in statistical terms, the commitment hypothesis may predict an interaction between the entropy at the point at which commitment can be made and the conditional probability of the eventual analysis.

In the case of Language F, for example, the uncertainty about the identity of the full word is lower when the first phoneme is [b] than when it is [t]. According to the commitment hypothesis, then, listeners will be more likely to activate full representations after [b] than after [t]. Although

the conditional probability of [a] being the second phoneme is equal in both cases, the commitment hypothesis would predict that [a] is more likely to trigger costly revision after the low-uncertainty [b] than after the high-uncertainty [t].

2.4 Estimation

2.4.1 Infinite languages

The examples so far have involved finite languages, where surprisal and entropy can be easily computed by enumerating all of the words in the language. Estimating these quantities is more complicated in the case of infinite languages characterized by a grammar. Efficient methods have been developed to estimate the conditional probability distribution over the next word induced by probabilistic context-free grammars and some related formalisms (Jelinek & Lafferty, 1991; Roark, 2001; Stolcke, 1995; Wu, Bachrach, Cardenas, & Schuler, 2010). Other studies have used neural networks, in which the probability distribution over the upcoming word can be read off the output layer of the network (Frank & Bod, 2011).

Grammar-based entropy is more difficult to estimate, and has more degrees of freedom. In most syntactic frameworks, the identity of the sentence is assumed to include both a sequence of words and a structural description. In ambiguous sentences such as Chomsky's example *flying planes can be dangerous*, then, some uncertainty about the identity of the sentence remains even after all of the input has been processed. Even word strings that ultimately identify a unique parse tend to have prefixes whose partial parses are ambiguous, e.g. (Marcus, 1980):

(3) a. Have the students take the exam today.

b. Have the students taken the exam today?

The first three words, *have the students*, are compatible with two parses: one where *the students* are the object of *have*, which is used as a transitive verb in the imperative (in the sense of "cause

the students to"), and another where *the students* are the subject and *have* is an auxiliary verb. This is not an unusual case: most sentences are at least temporarily ambiguous. In general, then, there is uncertainty about both the correct analysis of the prefix and the continuation of the sentence.

Typically, infinitely many sentences will be compatible with a given prefix $w_{1..k}$ (for example, one can always conjoin another clause at the end of the sentence). This makes it impossible to calculate entropy by simply enumerating the probabilities of all parses. For probabilistic context free grammars (PCFGs), a closed form formula exists for calculating the entropy of the distribution over sentences defined by a grammar; this includes both the correct analysis of the prefix and the continuation of the sentence (Grenander, 1967). To calculate the entropy of sentences that have a particular prefix, this formula can be applied to the PCFG resulting from the intersection of the original grammar with a finite state automaton that represents the prefix $w_{1..k}$ (Bar-Hillel, Perles, & Shamir, 1964; Nederhof & Satta, 2003; see Yun et al., 2015 for an overview of the procedure).

This procedure for calculating the entropy of a PCFG has been argued to be too computationally intensive for state-of-the-art PCFGs, which tend to have a large number of nonterminals (since they are either lexicalized or have a heavily split symbol set). In particular, this issue applies to the grammar used in the popular Roark (2001) parser, which has served as the source of surprisal estimates in most broad-coverage corpus research. Later work (Roark, 2011) has proposed to approximate the uncertainty about the identity of the sentence using the entropy of the set of derivations resulting from extending all the existing partial derivations by a single word, which can be any one of the words in the lexicon ("predictive entropy"). In other words, if $D(w_1 \dots w_i)$ is the set of partial derivations corresponding to the string of words $w_1 \dots w_i$, then predictive entropy is given by

$$H\left(\bigcup_{w \in A} D(w_1 ... w_k w)\right)$$

This calculation of predictive entropy is conceptually simple, though extending the current derivation with every possible word in the lexicon still requires significant computational effort. It is unknown whether extending the current derivation to 0 only a single upcoming word is a good approximation of full sentence entropy.

A related strategy is employed by Frank (2010, 2013) to calculate entropy in a recurrent neural network (RNN).[6] Specifically, the prefix $w_1, \ldots, w_n$ observed thus far is extended by all potential k-word-long continuations, and the entropy over the distribution of k-word-long continuations is taken as an approximation of the entropy over the identity of the sentence. Since there is a large number of potential k-word-long continuations, low-probability ones are pruned. Frank shows that entropy estimates based on $k = 4$ (the highest k he explores) are very different from those based on $k = 1$. On the other hand, entropy estimates based on $k = 4$ are fairly similar to those based on $k = 3$, tentatively suggesting that $k = 4$ is a reasonable approximation for full entropy ($k = \infty$). This measure is purely predictive; since RNNs only predict the next word and do not construct a syntactic analysis of the input, there is never any ambiguity about the past—all of the uncertainty is predictive.

Finally, Wu et al. (2010) derive entropy estimates from yet another model: a Hierarchical Hidden Markov Model (Schuler, AbdelRahman, Miller, & Schwartz, 2010). The estimate that Wu et al. (2010) propose quantifies *only* past uncertainty, that is, uncertainty about the correct structural analysis of the input so far, and has no predictive component at all. This estimate is therefore likely to differ substantially from full sentence entropy.

To summarize, PCFGs are the only generative model of sentences for which a known procedure exists for calculating sentential entropy. This procedure is computationally expensive, and has mostly been applied to small hand-created grammars (e.g., J. Hale, 2003b, 2006, though see Linzen & Jaeger, 2014 for an application to a partially lexicalized broad coverage grammar). Other studies have used approximations whose validity has not been experimentally established; some of these approximations may greatly underestimate sentential entropy.

[6] RNNs are connectionist sequence models that don't explicitly encode hierarchical structure (Elman, 1990).

Form	Count	P_{ML}	P_{smoothed}
schmooze	8	0.8	$^{9}/_{14} = 0.64$
schmoozes	0	0	$^{1}/_{14} = 0.07$
schmoozed	2	0.2	$^{3}/_{14} = 0.21$
schmoozing	0	0	$^{1}/_{14} = 0.07$
		$H = 0.72$	$H = 1.43$

(a) Infrequent verb

Form	Count	P_{ML}	P_{smoothed}
kick	750	0.75	0.748
kicks	20	0.02	0.021
kicked	200	0.2	0.2
kicking	30	0.03	0.031
		$H = 1.04$	$H = 1.05$

(b) Frequent verb

Table 2.8: The effect of smoothing on entropy: smoothing has a much stronger effect when there are cells with very low counts; in this example, *work* has a higher entropy paradigm before add-one smoothing, but a lower entropy paradigm after smoothing.

2.4.2 Smoothing

The probability of a linguistic unit is typically estimated from its relative frequency in a corpus. Suppose that we are trying to estimate the probability distribution of inflectional suffixes for a given English verb. The inflected forms of *kick*, for example, are *kick, kicks, kicked* and *kicking*. A simple way to estimate the probability of a suffix given the stem is to count how many times each inflected form occurs in a corpus and divide the outcome by the total number of times the verb occurred in any of its forms. For example, the probability of *kick* being followed by *-ed* would be

$$P(S = \textit{-ed} \,|\, \textit{v=kick}) \approx \frac{C(\textit{kicked})}{C(\textit{kick}) + C(\textit{kicks}) + C(\textit{kicked}) + C(\textit{kicking})}$$

where $C(w)$ denotes the number of times the form w occurred in the corpus. One drawback of this method is that an inflected form that happened not to appear in the corpus—due to data sparsity—will be assigned zero probability. This language model would therefore conclude that rare forms are ungrammatical—an undesirable conclusion.

Consider the inflectional paradigm for the infrequent verb *schmooze* in Table 2.8a. The entropy of this distribution is $H_{\text{MLE}} = 0.72$. This probability distribution makes the implausible prediction that *schmoozing* will never occur in any English text: a reader that came across this form in a text would be infinitely surprised. The standard solution to this problem is to "smooth" the distribution

by taking some of the probability mass assigned to the attested forms and spreading it across the unseen forms. In additive smoothing, for example, we pretend to have observed all outcomes of the distribution α more times than we actually did; Table 2.8a illustrates the application of this method for $\alpha = 1$ ("add-one smoothing"). After applying this smoothing method, we get $H_{\text{smoothed}} = 1.43$, almost double the entropy of the unsmoothed distribution.

Smoothing affects different information theoretic measures in different ways. The difference between smoothed and unsmoothed surprisal estimates will be larger in distributions with a low total count (such as the suffix distribution for *schmooze*). This difference will be particularly acute for low probability outcomes within those distributions (*schmoozing*), and less so for higher probability outcomes (*schmooze*). In the case of entropy and entropy reduction, the effect may be even more dramatic: suffix distributions for low-frequency verbs with many zero-probability outcomes will have artificially low entropy, leading to a strong correlation between the frequency of the verb and its entropy (Table 2.8). This issue will be more severe the more potential outcomes there are. For example, suppose we are estimating the entropy of the distribution of nouns that can follow different English adjectives. In this scenario, a frequent adjective such as *large* would occur in a corpos with thousands of different nouns, whereas an infrequent one such as *humongous* would occur with many fewer nouns. Even though *humongous* may be compatible with the same nouns as *large*, then, its entropy will be much lower.

Despite the correlation between unsmoothed entropy estimates and the total frequency of the paradigm, smoothing has not received much attention in most empirical work that used entropy, in particular in morphological processing. This would be an important issue to explore in future work. Simple additive smoothing is unlikely to be an optimal technique (see Goodman, 2001; Manning & Schütze, 1999 for a review), and smoothing methods specific to entropy may provide better results (Valiant & Valiant, 2013). Moreover, different linguistic phenomena may call for different smoothing methods. In the case of morphological paradigms, for example, it may make sense to smooth based on the inflectional paradigm of a "typical" verb in the language, constructed by averaging across all verbal paradigms.

2.5 Empirical findings

2.5.1 Sentence processing

A wide range of studies have documented surprisal effects on reading times. These studies have used a variety of language models, including trigram models (Smith & Levy, 2013), neural networks (Frank, 2013) and probabilistic grammars (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; Fossum & Levy, 2012; Roark, Bachrach, Cardenas, & Pallier, 2009; Wu et al., 2010). A recent study has shown that surprisal is also a useful predictor of the magnitude of the N400 ERP component (Frank, Otten, Galli, & Vigliocco, 2015). The studies that are framed explicitly in terms of surprisal join a long tradition of behavioral and neural studies that have demonstrated that words that have a higher conditional probability are read faster and elicit attenuated neural responses compared to lower probability words (DeLong, Urbach, Groppe, & Kutas, 2011; Ehrlich & Rayner, 1981; Kutas, DeLong, & Smith, 2011; S. McDonald & Shillcock, 2003b), though surprisal predicts a logarithmic rather than linear relationship between conditional probability and the dependent measure.

The empirical picture regarding entropy and entropy reduction effects is more complicated. A variety of grammatical formalisms and approximations have been used to calculate these quantities (see Section 2.4.1). The majority of studies have focused on the entropy reduction hypothesis; even those studies, however, have evaluated it in subtly different ways.

J. Hale (2003b, 2003a, 2006) used hand-constructed PCFGs and Minimalist Grammars to show that the entropy reduction hypothesis predicts certain processing difficulty contrasts from the sentence processing literature. In the classic garden path sentence *the horse raced past the barn fell*, for instance, the amount of entropy reduction at *fell* is very high. According to Hale, that explains the processing difficulty that readers experience at that word. Many garden path effects are predicted by surprisal as well, however; in particular, J. Hale (2001) presents evidence that the surprisal of the disambiguating word *fell* is also high. Since surprisal isn't controlled for in these simulations, these cases do not necessarily constitute strong evidence for the entropy reduction

hypothesis. One contrast that has been argued to be accounted for by entropy reduction but not surprisal is the asymmetry between subject and object relative clauses in Asian languages (Yun et al., 2015).

Other studies have reported effects of entropy reduction controlling for surprisal. Linzen and Jaeger (2014) report an effect of entropy reduction derived from a PCFG in a controlled self-paced reading experiment. Frank (2013) reports an effect of entropy reduction in the expected direction in a self-paced reading corpus; entropy in this study was calculated from the set of four predicted words in a Recurrent Neural Network. Likewise, Wu et al. (2010) found an effect of entropy reduction in the expected direction, again in a self-paced reading corpus. This effect held for closed-class words only. As noted in Section 2.4.1, entropy in this study was of the set of partial Hierarchical HMM derivations consistent with the string so far (i.e., no predictive component).

By contrast with the studies mentioned so far, Roark et al. (2009) tested for an effect of entropy rather than entropy reduction; they therefore implicitly evaluated the competition hypothesis. As noted in Section 2.4.1, their measure ("predictive entropy") captures the uncertainty about all partial parses of the sentence up to the current word, extended to include the part-of-speech of a single predicted word. This measure was evaluated on a self-paced reading corpus. They found that predictive entropy at a word was a relatively strong predictor of reading times on the same word (in particular, it was a stronger predictor than word frequency).

2.5.2 Visual word recognition

This section discusses applications of information theoretic variables for predicting dependent measures associated with the recognition of visually presented individual words. The main dependent measure in this domain is lexical decision reaction times.

2.5.2.1 Frequency as informativity

The inversely logarithmic relationship between word frequency and reaction times in word recognition tasks in one of the most established findings in psycholinguistics (Howes & Solomon, 1951):

if $F(w)$ is the frequency of w, then

$$\text{RT}(w) \propto -\log_2 F(w) \tag{2.19}$$

This finding can be interpreted as a linear relationship between reaction times and word surprisal (Moscoso del Prado Martín et al., 2004; Milin, Kuperman, Kostic, & Baayen, 2009). This is a special case of the general effect of surprisal in sequential processing: in the general case, surprisal predicts that

$$\text{RT}(w_n) \propto -\log_2 P(w_n|w_{1..n-1})$$

This quantity reduces to $-\log_2 P(w_n)$ when there is no context (Norris, 2006).

Much of the interest in information theoretic measures in isolated word recognition has centered on morphologically complex words. These words consist of multiple morphemes, each having a distinct semantic or functional role. For example, *touched* consists of *touch* and *-ed*. Each of the morphemes can be seen as a separate source of information about the identity of the complex word. Several lexical decision studies have found that reaction times to complex words are affected by the frequency—or, equivalently, informativity—of the stem (base), e.g., *touch* in *touched* (Taft, 1979; Ford, Davis, & Marslen-Wilson, 2010). There is less evidence for an effect of the frequency of affixes such as *-er* on lexical decision reaction times, though both affix frequency and the conditional probability of the affix given the stem affect the M170 neural responses to complex words (Solomyak & Marantz, 2009a).

Kostić and colleagues argue that the reason for the absence of affix frequency effects is that the information conveyed by a functional suffix needs to be normalized by the number of functions that this suffix could have: an inflected form with multiple semantic or syntactic functions is taken to be more informative (Kostić, 1991, 1995; Kostić, Marković, & Baucal, 2003). These studied analyzed

[6]There is some debate as to whether the relationship is in fact logarithmic (Murray & Forster, 2004; Adelman & Brown, 2008); this issue falls outside the scope of this review.

lexical decision reaction times to Serbian inflected words, such as *konj* 'horse (nominative)' and *konja* 'horse (genitive)'. Specifically, if F_a is the total frequency of an affix such as nominative or genitive (summed across all stems), and N_a is the number of syntactic functions filled by the affix, then the "information load" of a as defined by Kostić and colleagues is

$$I(a) = -\log_2 \frac{F_a/N_a}{\sum_{a'} F_{a'}/N_{a'}} \tag{2.20}$$

For example, although the nominative plural form *konji* and the accusative plural form *konje* have similar probability, the former has three functions ("subject", "predicate", "in exclamations") and the latter has 58 functions (e.g., "top down movement", "price" and "putting together"). Given the difference in number of functions between the two cases, Kostić's model predicts longer lexical decision latencies in response to the accusative form.

2.5.2.2 Morphological family and competition

The same stem can typically combine with multiple different affixes. A word's **morphological family** is defined as the set of words that are morphologically related to the word in question. The family can be inflectional (e.g., the inflectional family of *think* includes words such as *thinks* and *thinking*), derivational (*thinker*, *rethink*) or both. Monomorphemic words with larger morphological families are recognized faster in visual lexical decision (Baayen et al., 1997; Moscoso del Prado Martín et al., 2005; Schreuder & Baayen, 1997) and auditory lexical decision (Balling & Baayen, 2008). The size of a word's morphological family also affects the latency of the M350 component in MEG (Pylkkänen, Feintuch, Hopkins, & Marantz, 2004) as well as gaze durations in natural reading (Juhasz & Berkowitz, 2011; Kuperman, Bertram, & Baayen, 2008). These findings have been interpreted as indicating that a word's morphological family is activated when the word is encountered.

Following those studies, Moscoso del Prado Martín et al. (2004) proposed that family size should be replaced with the entropy of the probability distribution defined by the morphological

family. In their study, higher morphological entropy led to shorter lexical decision reaction times. This suggests that increased competition between the inflected or derived forms constitutes a cue to the lexicality of the form and speeds up the lexical decision (see the discussion of the competition hypothesis in Section 2.3.3). The task-specific interpretation of this result is supported by the finding that this variable does not affect naming latencies (Baayen, Feldman, & Schreuder, 2006).

Semantically related and unrelated derived forms appear to have opposite effects on the recognition of the stem: semantically related derived forms tend to shorten reaction times, whereas unrelated derived forms (e.g., *department* for *depart*) interfere with recognition (Moscoso del Prado Martín et al., 2005). This suggests that some kinds of competition are beneficial to word recognition while others can be harmful (cf. Q. Chen & Mirman, 2012).

2.5.3 Spoken word recognition

Whereas in visual word recognition the entire word is perceived at once, spoken words unfold sequentially in time. The sequential information update framework discussed in Section 2.2 is fully compatible with the cohort model of spoken word recognition. In this model, all of the words that are compatible with the sounds heard so far are activated; the set of the words that are activated at any given point is referred to as the *cohort* (Marslen-Wilson, 1987). Each phoneme can therefore be associated with its surprisal and with the entropy of the cohort after the phoneme has been perceived. Standard behavioral paradigms such as auditory lexical decision are not well suited to studying phoneme-specific predictors: they yield a single dependent measure for the entire word, and it is not obvious how the informativity of each phoneme should be related to that single number. We begin by discussing attempts to deal with this issue in behavioral studies. We then survey studies that used neural measures, which can track the response to each individual phoneme.

2.5.3.1 Reaction time studies

Cohort entropy was used as a predictor in a series of auditory lexical decision experiments that investigated morphological family effects on reaction times in spoken word recognition. Kemps, Wurm, Ernestus, Schreuder, and Baayen (2005) presented participants with monomorphemic English and Dutch words. Cohort entropy was calculated at the final segment of the stem. The cohort included both morphologically related and unrelated words; for example, the cohort for *bake* included *baking*, *bakery* and also *bacon*. Kemps et al. (2005) found that higher word-final cohort entropy was associated with longer lexical decision reaction times.

In a lexical decision experiment on prefixed Dutch words (e.g. *be-vecht* 'fight') and matched monomorphemic words, Wurm, Ernestus, Schreuder, and Baayen (2006) explored how reaction times were affected by cohort entropy at three different points in the word: two phonemes into the words, in the middle of the word,[7] and at the end of the word. They found that higher entropy early in the word correlated with shorter lexical decision RTs; higher entropy in the middle of the word correlated with longer RTs, regardless of whether or not it was prefixed; and higher entropy at end of the word again correlated with shorter RTs. Their interpretation of the results is specific to the lexical decision task: if the beginning of the item is consistent with many words in the language, the participant will be more likely to think it is a word. At the end of the word, the cohort consists mostly of morphologically related words, which again contribute to the impression of lexicality (higher morphological family size was also correlated with shorter RTs). In the middle of the word, higher uncertainty means that the participant still does not know which individual lexical root it is, and is therefore slowed down.

[7]More specifically, at the "CRUP", or Conditional Root Uniqueness Point. CRUPs are only relevant for prefixed words like *discredit*; the CRUP identifies the point in a free root (*credit*) at which the identity of the free root is unique given the prefix (*dis-*); in this case, it's the *r*. This is different from the uniqueness point, because *discredit* diverges later from words like *discrepancy*, in which the material following *dis-* is not a free morpheme.

Finally, Balling and Baayen (2012) focus on the processing cost of updating the probability distribution over the cohort after each phoneme, as measured by the relative entropy between the distributions before and after the phoneme. As we have pointed out, this quantity is equivalent to the surprisal of the incoming phoneme. They summarize the vector of phoneme surprisals using the slope of the cumulative surprisal function, i.e., the sum of the surprisals of all of the phonemes so far. Formally, the cumulative surprisal at the k-th phoneme is:

$$c_k = -\sum_{i \leq k} \log_2 P(\phi_i | \phi_1, \ldots, \phi_{i-1})$$

They then find the best-fit linear function for $c_1, \ldots, c_n$—i.e., predicting the cumulative surprisal after the k-th phoneme from its position k—and use this slope to predict lexical decision RTs. This slope is in general higher when most of the surprisal is concentrated in the earlier phonemes and lower when it is concentrated in later phonemes. They report a complex pattern of results: early in the experiment participants are slowest on words where surprisal is higher towards the end of the word (shallow surprisal slopes); as the experiment progresses, the pattern reverses and participants become slowest on words whose *earlier* phonemes are surprising (steep surprisal slope).[8]

2.5.3.2 Neural studies

The complicated empirical picture that emerges from the lexical decision studies discussed in this section shows illustrates the difficulties in linking phoneme surprisal and cohort entropy at different points in the word to lexical decision RTs; progress in this area would likely require a more

[8] It is worth pointing out that cumulative surprisal c_k is equivalent to the surprisal of the prefix $\phi_1 \phi_2 \ldots \phi_k$, due to the chain rule:

$$-\log_2 P(\phi_1 \ldots \phi_k) = -\log_2 (P(\phi_1) P(\phi_2|\phi_1) \cdots P(\phi_k|\phi_1, \ldots, \phi_{k-1})) =$$

$$-(\log_2 P(\phi_1) + \log_2 P(\phi_2|\phi_1) + \ldots + \log_2 P(\phi_k|\phi_1, \ldots, \phi_{k-1})) = c_k$$

precise model of the auditory lexical decision task. The relationship between the sequential informativity paradigm and continuously recorded neural signals is more straightforward. Gagnepain et al. (2012) used MEG to measured participants' neural activity while the participants were listening to isolated words. They compared the success of two variables in in predicting neural activity: "prediction error", which was linear in the conditional probability, and the entropy of the cohort. They found that prediction error affected neural response but cohort entropy did not. The authors interpret their result as arguing for phoneme prediction and against competition among activated lexical items.

Ettinger et al. (2014) had participants perform a lexical decision task during an MEG recording. They tested whether the magnitude of neural activity in left auditory cortex can be predicted from phoneme surprisal and cohort entropy, under the assumption that these variables would affect neural activity 150 ms after the onset of the relevant sounds were played. They found that higher surprisal led to increased neural activity, in line with the findings of Gagnepain et al. (2012). This effect was only observed towards the end of the word. The effect of cohort entropy had a complicated shape. In the first half of the word, higher cohort entropy was associated with lower neural activity; in the second half of the word, the direction of the correlation flipped, and higher entropy was associated with increased activity. Ettinger et al. (2014) interpret this pattern in terms of the commitment hypothesis (Section 2.3.4): the cohort is more strongly activated when it is smaller; if the cohort is large, more phonemes need to be accumulated before words are activated.

In a recent auditory lexical decision study in English with concurrent MEG recording, Brennan, Lignos, Embick, and Roberts (2014) found that cohort entropy computed after the first phoneme of a word correlated with slower reaction times—the opposite of the findings of Wurm et al. (2006). They were unable to find a neural correlate of this reaction time effect in a spectral (frequency-band based) analysis.

In summary, electromagnetic recordings appear to be a promising technique to study phoneme-by-phoneme informativity and cohort activation measures; since this field is still in its infancy, however, a unified empirical picture has yet to emerge.

2.6 General discussion

This review has surveyed the ways in which the information theoretic measures of surprisal, entropy and relative entropy have been used in the study of human language comprehension. These measures have been used in three areas: sequential phoneme perception in spoken word recognition, visual word recognition (which is not sequential) and parsing in sentence processing. We will now provide a general comparative overview of the use of these measures across the three areas and point out directions for future research.

Spoken word recognition and parsing both involve sequential updating of beliefs over an underlying representation whose identity is gradually revealed. In both of those scenarios it is therefore natural to apply the two measures of information gain discussed in Section 2.2: surprisal and entropy reduction. Surprisal has been explored extensively in sentence processing, and there is convincing evidence that it affects reading times and neural activity. Phoneme surprisal has recently begun to be studied in spoken word recognition research as well, though the technical challenges in obtaining appropriate dependent measures have made progress in this direction slower.

A series of studies have searched for potential effects of entropy reduction in sentence processing, some with positive results (Frank, 2013; Linzen & Jaeger, 2014; Wu et al., 2010; Yun et al., 2015). Entropy reduction over full sentences is in general difficult to calculate. Consequently, the empirical picture is complicated by the fact that studies to date have used a range of grammatical formalisms and approximations to sentential entropy. A systematic comparison across those approximations would be helpful in advancing this area of research.

To my knowledge, there are no studies of entropy reduction in spoken word recognition. This is perhaps unsurprising giving the small number of studies that have used phoneme-by-phoneme predictors. At the same time, calculating entropy over a finite lexicon is dramatically easier than calculating entropy over an infinite set of sentences compatible with a prefix; the methodological issues that beset entropy reduction research in sentence processing do not arise in spoken word recognition. This makes this area a potentially promising testing ground for this hypothesis.

Conversely, the competition hypothesis has had more impact in spoken word recognition than in sentence processing. A common assumption in the spoken word recognition field, going back to the cohort model, is that all of the words in the lexicon that are compatible with the prefix are activated. This has led researchers to look for effects of the entropy of the cohort on reaction times and neural activity. In contrast with surprisal and entropy reduction, the motivation for using entropy in this context is typically not related to information: entropy is seen as a tool for quantifying the total degree of activation, taking into account both the number of cohort members and their relative probabilities.

Entropy has been used in word recognition research more generally to quantify the degree of predicted activation of various sets of words that are related to word being recognized. Depending on the relationship between the word being recognized and the set of related words in question (cohort, morphological family, etc), as well as on the task the participants are performing, greater activation of related words may either facilitate or inhibit word recognition (Moscoso del Prado Martín et al., 2004; Simon et al., 2012).

In future research, it may be worth examining whether entropy is the optimal way to quantify support from or competition with related words. It could well be the case that a single strong competitor inhibits recognition more than nine competitors with a probability of 0.1 each, despite the fact the entropy is higher in the latter case. In contrast to the information theoretic motivation for expecting entropy reduction or surprisal effects, no computational-level motivation has been proposed for an entropy effect in word recognition; this makes it even more pressing to implement a mechanistic model of word recognition and derive predictions from it (Q. Chen & Mirman, 2012).

While the common assumption in word recognition research is that high entropy of the distribution of related words (cohort, family) leads to increased neural activation (which, as mentioned above, can either lead to quicker or to slower recognition), this is not the only possible scenario. The commitment hypothesis argues that in some cases processing is delayed when uncertainty is high (Ettinger et al., 2014; Linzen, Marantz, & Pylkkänen, 2013). For example, if participants are

recognizing a word that is unambiguously a noun, they may activate the syntactic representation of the noun category; conversely, if a word is ambiguous between multiple syntactic categories, they may leave the syntactic category unspecified, and only activate it in a disambiguating context.

3

Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions

3.1 Introduction[1]

One of the major challenges that readers face when processing a sentence is inferring its syntactic structure (parsing it). There is growing evidence that people parse sentences in an incremental and predictive fashion: Each incoming word is used to revise existing hypotheses about the correct parse of the sentence and predict upcoming syntactic structure (Altmann & Kamide, 1999; J. Hale, 2001; Levy, 2008a). These predictions are probabilistic: There is a continuous relationship between predictability and processing difficulty (Boston et al., 2008; Demberg & Keller, 2008; Jennings, Randall, & Tyler, 1997; S. McDonald & Shillcock, 2003a, 2003b). This likely reflects a strategy whereby readers prepare to process upcoming linguistic material in proportion to its probability (DeLong et al., 2005; Smith & Levy, 2013).

[1] This work was done in collaboration with Florian Jaeger, and is now in press in *Cognitive Science*. A previous version of this work was presented as Linzen and Jaeger (2014).

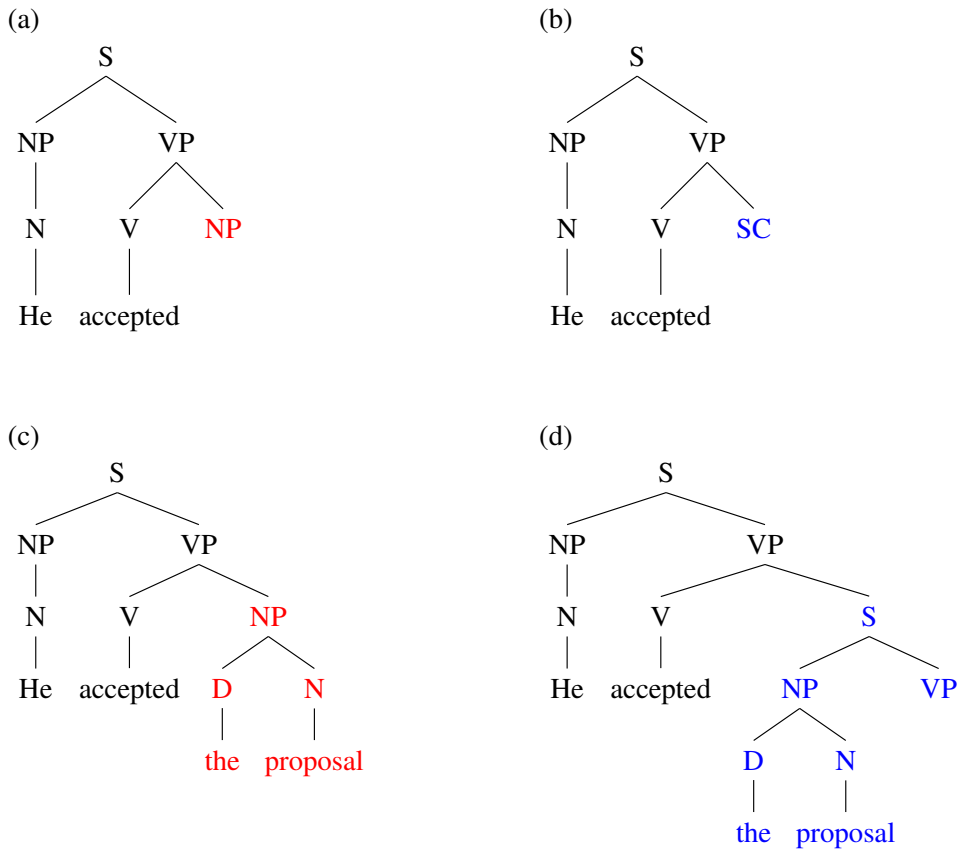


Figure 3.1: Incremental parses with single-step prediction while reading the sentence he *accepted the proposal was wrong*. (a) and (b) represent parses after the verb has been read and its complement type has been predicted. (c) and (d) represent the parses after the ambiguous region *the proposal* has been incorporated into (a) and (b) respectively.

As an example, consider the sentence in (1):

(4) He accepted the proposal was wrong.

The verb *accept* can be followed by two types of complements, or *subcategorization frames*: a noun phrase (as in *accept a gift*) or a sentential complement (as in *accept that you've lost*). In actual usage, *accept* occurs much more frequently with the noun phrase (NP) frame than with the sentential complement (SC) frame. Having read the word *accepted*, then, the reader can form a

strong prediction for an NP, and possibly a weaker prediction for an SC (Fig. 1a and 1b). The next words, *the proposal*, are compatible with both parses: They can either serve as the verb's direct object or as the subject of an SC (Fig. 1c and 1d). Finally, the words *was wrong* disambiguate the sentence in favor of the low-probability SC parse. In line with the predictive parsing hypothesis, the disambiguating region *was wrong* tends to be read more slowly when the verb favors the NP frame (e.g., *accept*) than when it favors the SC frame (e.g., *prove*) (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell et al., 1993).

While there is ample evidence that expectation violation, as in the example just discussed, can lead to processing difficulty, less is known about the generation and maintenance of those expectations. Suppose, for example, that the verb *accepted* in sentence (1) were replaced by *forgot*. The verb *forgot* is similar to *accept* in that it is biased against an SC continuation, but differs from it in that it has a more diverse set of potential complements. Specifically, in addition to an NP (*forgot my birthday*, 55%) and an SC (*forgot he was supposed to go*, 9%), this verb can be followed by a prepositional phrase (*forgot about the party*, 18%) or an infinitive (*forgot to buy groceries*, 14%). Consequently, there is a greater degree of uncertainty about upcoming syntactic structure after *forget* than after *accept*. Does this difference between *forget* and *accept* affect processing difficulty, and if so, how?

Following standard practice, we quantify uncertainty about a probabilistic outcome using the Shannon entropy of the distribution:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3.1}$$

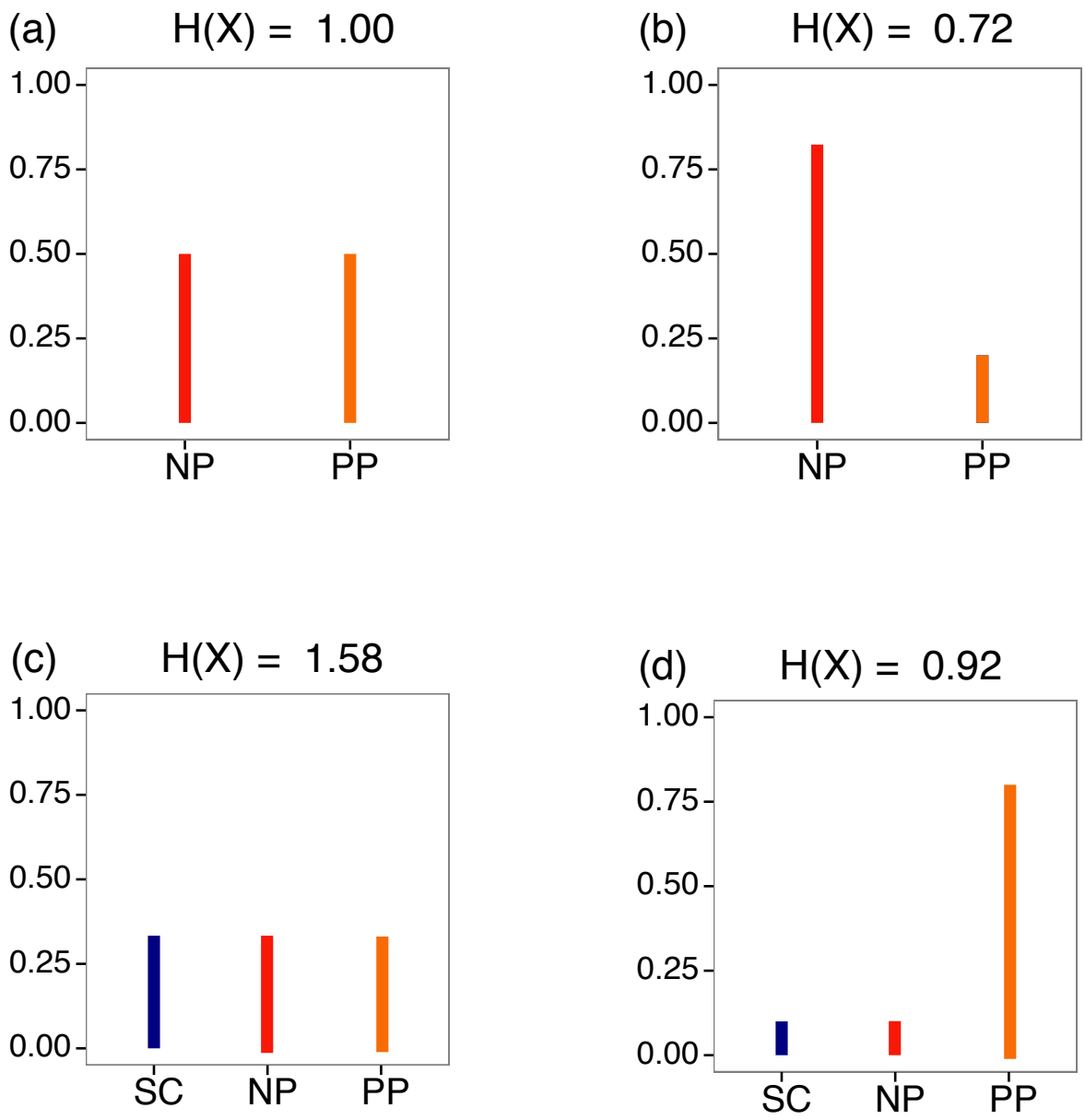


Figure 3.2: Entropy values for four examples of subcategorization distributions: (a) two balanced frames; (b) two unbalanced frames; (c) three balanced frames; (d) three unbalanced frames. When probability is evenly distributed across subcategorization frames, verbs with more frames have higher entropy (compare (a) to (c)). For the same number of frames, entropy is lower the less balanced the distribution (compare (a) to (b), or (c) to (d)).

Entropy is higher the more potential parse completions there are and the more uniformly distributed the probabilities of those parses are. For instance, when there is only one parse completion, the entropy is 0 bits. With two equiprobable completions the entropy is 1 bit, and with three

equiprobable options the entropy is 1.58 bits. If one of the three options is much more likely than the others, the entropy can be lower than the entropy of two equiprobable options (see Fig. 3.2).

We focus on two ways in which readers' uncertainty about syntactic expectations, as quantified by entropy, may affect processing difficulty. First, it may be costly to generate and maintain a larger number of predictions that compete with each other, especially if their probabilities are similar. We term this hypothesis the competition hypothesis (Elman, Hare, & McRae, 2005; McRae et al., 1998). A second hypothesis, the entropy reduction hypothesis, proposes that it is reduction in uncertainty that is costly rather than the mere existence of uncertainty (J. Hale, 2006; Yun et al., 2015). Under this hypothesis, an increase in uncertainty does not affect processing; that is, if H_i the entropy at the i-th word, then entropy reduction at the i-th word is given by $\max\{H_i - H_{i-1}, 0\}$.

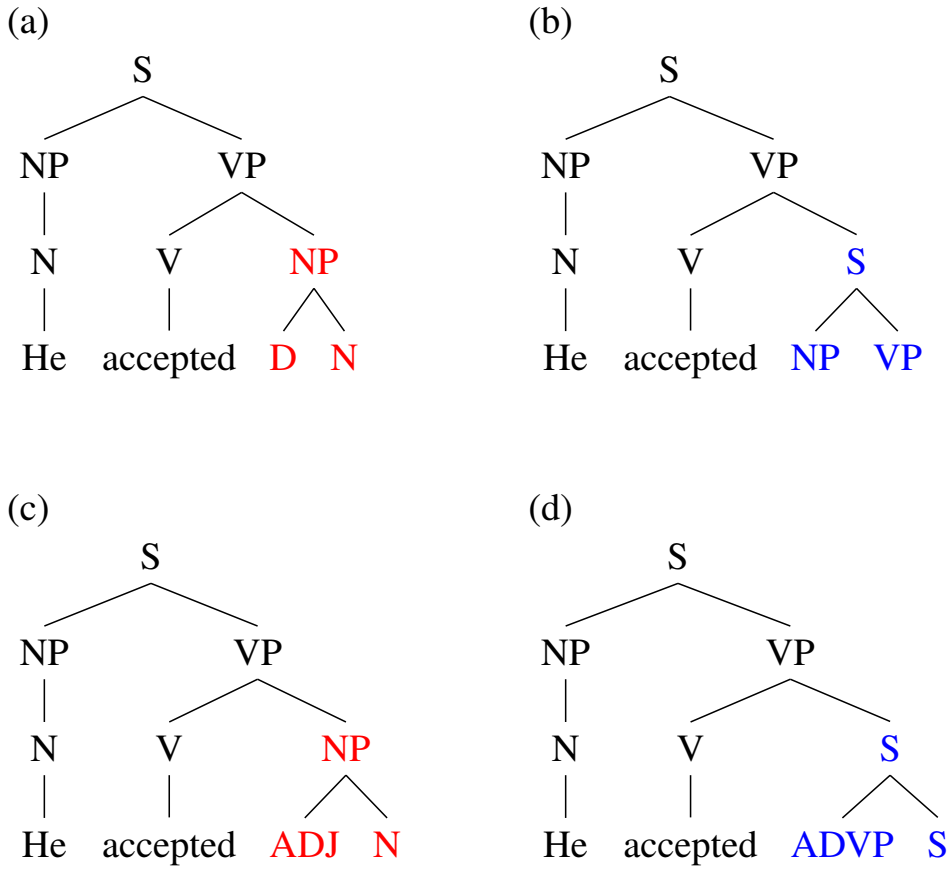


Figure 3.3: Predicting syntactic structure an unlimited number of derivation steps ahead: In addition to predicting the category of the complement (NP or SC) the reader predicts its internal structure. (a) prediction of a simple determiner + noun phrase (*he accepted the present*); (b) prediction of a noun phrase with an adjective (*he accepted the nice present*); (c) prediction of an SC with an intransitive verb (*he accepted that he lost*); (d) prediction of an SC with a transitive verb (*he accepted that he lost her*).

The examples discussed so far have focused on the expectations that comprehenders generate immediately after processing a verb based on its subcategorization frequencies – in other words, expectations for the next immediate node in the parse tree following the verb (Fig. 1). Yet it is possible that readers generate more detailed syntactic predictions, consisting of multiple derivation steps. For instance, instead of simply predicting an NP, they might probabilistically predict both an NP consisting of a determiner and a noun (*the present*) and an NP consisting of a determiner,

an adjective and a noun (*the nice present*) (Fig. 3.3). The depth of syntactic structure predicted by readers is an open question. We investigate two endpoints of the prediction depth spectrum – prediction of the next syntactic step of the derivation (single-step prediction) and prediction of the entire syntactic structure of the sentence (full prediction).

A number of recent studies have begun to explore the effect of uncertainty on reading times (Frank, 2013; Frank et al., 2015; J. Hale, 2003b, 2006; Roark et al., 2009; Wu et al., 2010; Yun et al., 2015) and neural measures (Frank et al., 2015), some with positive results. However, as we detail next, the conclusions that can be drawn from these studies are limited by the considerable variability between them, as well as lack of critical controls. This state of affairs has led others to question the extent to which these studies provide support for the role of uncertainty-based hypothesis, in particular the entropy reduction hypothesis (Levy & Gibson, 2013). We begin by summarizing the previous studies; we then present the motivation for the current study.

3.1.1 Previous work on entropy in sentence processing

One class of studies has demonstrated that the entropy reduction hypothesis can predict qualitative findings from the sentence processing literature, such as the processing difficulty at the disambiguation point in garden path sentences (*the horse raced past the barn fell*; J. Hale, 2003b) or the asymmetry between object and subject relative clauses (Yun et al., 2015). These predictions have not been evaluated on reading times, and have not controlled for alternative sources for the difference in processing difficulty between the constructions, such as surprisal (J. Hale, 2001), memory cost (Grodner & Gibson, 2005) or similarity based interference (Lewis & Vasishth, 2005).

Other studies have assessed the effect of entropy in reading times corpora (Frank, 2013; Roark et al., 2009; Wu et al., 2010). Roark et al. (2009) found a positive effect of entropy over syntactic parses, supporting the competition hypothesis. Support for the entropy reduction hypothesis comes from two studies: Wu et al. (2010) found that entropy reduction had a positive effect on reading times, though only for closed class words, and Frank (2013) found a positive effect of entropy reduction, for all words.

Several factors limit the information that can be gained from these studies. Foremost, most previous studies did not directly compare the competition and entropy reduction hypothesis. This is problematic since entropy and entropy reduction effects will tend to be highly correlated (this is confirmed in the present study). A comparison across these previous studies is further made difficult by methodological differences between these studies. First, previous work has used a wide variety of syntactic models, ranging from connectionist networks (Frank, 2013) through unlexicalized (J. Hale, 2003b) and lexicalized probabilistic context-free grammars (Roark et al., 2009) to Hierarchical Hidden Markov Models (Wu et al., 2010). It is thus unclear to what extent the divergent results of previous studies are due to differences in representational assumptions.

Second, the studies differed substantially in the way in which entropy (and thus entropy reduction) was calculated. For example, J. Hale (2003b) calculated entropy over full sentences (full entropy): this measure includes uncertainty both about the analysis of the part of the sentence read so far and about the rest of its structure. This contrasts with the entropy measure employed by Frank (2013), which captures uncertainty about a few upcoming words (that study used a neural network that does not represent syntactic ambiguity and therefore never has any uncertainty about the correct analysis of the past). The approach taken by Roark et al. (2009) falls between these two extremes in that it captures uncertainty about the analysis of the part of the sentence read so far, as well as uncertainty about the syntactic structure that would need to be constructed to accommodate a single upcoming word.

Third, some authors evaluated the entropy reduction hypothesis while assuming that increases in entropy facilitate processing (Frank, 2013), contrary to the formulation of the entropy reduction hypothesis (J. Hale, 2006). In some cases, effects of entropy reduction and of entropy may differ only in their sign (see below), making it impossible to distinguish between the competition and entropy reduction hypothesis as formulated by (J. Hale, 2006).

Finally, most previous studies examined processing difficulty across different types of structures, potentially confounding structural differences (and, e.g., correlated differences in memory cost, cf. Grodner & Gibson, 2005) with differences in entropy. Taken together, these differences

limit the conclusions that can be drawn from previous work about the role of uncertainty during language processing.

3.1.2 Contribution of the paper

As mentioned above, some researchers have questioned the viability of entropy reduction as a predictor of processing difficult, in part because of the complicating factors listed above (Levy & Gibson, 2013). The goal of the current study is to assess the effects of entropy and entropy reduction in the same materials, within the same syntactic framework, while avoiding structural confounds.

Following Roark et al. (2009), we evaluate the predictions of uncertainty-related hypotheses on human reading times. Following J. Hale (2003a), we undertake a detailed experimental and computational analysis of a specific class of sentences, using a simple syntactic framework: a probabilistic context-free grammar (PCFG) based on the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). We emphasize two distinctions that are not always clearly highlighted in the literature. First, we distinguish entropy effects from entropy reduction effects. Positive correlations between entropy and reading times are predicted by the competition hypothesis, whereas positive correlations between entropy reduction and reading times are predicted by the entropy reduction hypothesis (cf. J. Hale, 2006).

Second, we compare uncertainty in single-step prediction and uncertainty in full prediction. The entropy reduction hypothesis (as formulated in J. Hale, 2006) predicts only that the latter should be correlated with reading times. This distinction is thus critical, but has so far received little attention in the literature, which has employed a variety of different measures of entropy and entropy reduction under the implicit assumption that they are interchangeable. The results of the current study show that this assumption is wrong: Uncertainty in single-step prediction and uncertainty in full prediction can differ dramatically and exhibit qualitatively different correlations with reading times.

To avoid the potential confounds associated with comparing reading times across constructions, we manipulate syntactic uncertainty while keeping constant syntactic structure and associated confounding factors such as memory cost. We do so by varying the syntactic expectations induced by specific lexical items, as in the case of accept compared to forget above. Much previous work has shown that expectations in human language processing are sensitive to lexical information (Garnsey et al., 1997; Linzen et al., 2013; McRae et al., 1998; Trueswell et al., 1993).

We first describe a reading time experiment in which we vary single-step entropy by comparing verbs with different subcategorization distributions. We then discuss the relationship between single-step and full entropy, and assess how well each of these measures predict reading times.

3.2 Reading time experiment

The design of the experiment was modeled after Garnsey et al. (1997). Half of the sentences read by a given participant included the complementizer that, as in (2a) (henceforth referred to as unambiguous sentences), and half did not, as in (2b) (henceforth ambiguous sentences). We refer to this factor as Ambiguity.

(5) a. The men discovered that the island had been invaded by the enemy.
subject verb that ambiguous disambiguating rest

b. The men discovered the island had been invaded by the enemy.
subject verb ambiguous disambiguating rest

For consistency with previous studies, we refer to the subject of the embedded clause (*the island*) as the ambiguous region even when the sentence is unambiguous; likewise, we refer to the verbal complex *has been invaded* as the disambiguating region in both sentence types. In addition to the within-item ambiguity manipulation, two factors were manipulated between items: the subcategorization entropy of the main verb (high vs. low) and the surprisal of an SC given the verb (high vs. low). Subcategorization frequencies were taken from Gahl, Jurafsky, and Roland (2004) database (described in more detail below). We quantified SC bias using the surprisal (inverse log

probability) of an SC given the verb rather than raw conditional probability, based on evidence that the relationship between conditional probability and processing difficulty is logarithmic (Smith & Levy, 2013).

To factorially cross subcategorization entropy and surprisal, we leveraged the fact that many verbs occur with frames other than SC and NP.[2] For example, *find* and *propose* have a similar SC subcategorization probability (0.22 and 0.25 respectively) and thus similar SC surprisal. But *propose* occurs with multiple other frames (NP: 0.57; infinitives: 0.14), whereas for *find* the NP frame is the only alternative to SC that occurs with a substantial probability (0.72). As a result, *propose* has higher subcategorization entropy than *find* (1.56 vs. 1.09).

3.2.1 Predictions

In summary, the reading time study used a $2 \times 2 \times 2$ design: Ambiguity $\times$ subcategorization entropy $\times$ SC surprisal. If increased uncertainty causes a processing slowdown, as argued by the competition hypothesis, reading times at the verb region will be longer in the high subcategorization entropy conditions. Conversely, if it is reduction in uncertainty that causes processing slowdown, as argued by the entropy reduction hypothesis, we expect longer reading times on verbs with lower subcategorization entropy: Since entropy before the verb is matched across conditions (see below), verbs with lower subcategorization entropy reduce uncertainty more than verbs with higher subcategorization entropy. Finally, surprisal (J. Hale, 2001) predicts that disambiguation in favor of an SC parse will be more costly for high SC surprisal verbs than for low SC surprisal verbs (Garnsey et al., 1997). This should only occur in ambiguous sentences.

[2] If the subcategorization distribution has only two potential outcomes, SC and another frame (e.g., a direct object), then the surprisal of an SC, given by $-\log_2 p_{SC}$, is deterministically related to the entropy of the distribution, given by $-(p_{SC} \log_2 p_{SC} + (1p_{SC}) \log_2 (1p_{SC}))$.

3.2.2 Method

3.2.2.1 Participants

A total of 128 participants were recruited through Amazon Mechanical Turk and were paid $1.75 for their participation. Participants took 17 minutes on average to complete the experiment (standard deviation: 4.1 minutes).

3.2.2.2 Materials

We selected 32 verbs, eight in each of the cells of the 2×2 design defined by subcategorization entropy and SC surprisal. Subcategorization frequencies were obtained from the database of Gahl et al. (2004), which is based on the 18 million words of text comprising the Touchstone Applied Science Associates corpus (Zeno, Ivens, Millard, & Duvvuri, 1995) and the Brown corpus (Kucera & Francis, 1982). Gahl et al. classify subcategorization frames into six categories: transitive (*Klaus adore cookies*), intransitive (*we watched attentively*), quote (*he said "that's fine by me"*), finite sentential complement (*Trent yelled (that) the road was in sight*), infinitival complement (*she wanted to share her insight with others*) and "other".[3] Verbs were matched across conditions for their frequency and length. Frequency norms were obtained from the SUBTLEX-US corpus (Brysbaert & New, 2009). Table 3.1 shows the mean values and standard deviations across conditions for log-transformed verb frequency, subcategorization entropy and SC surprisal.

In the next step, 32 sentences pairs were created, one for each verb (a list of all items is provided in Appendix A). Each pair contained one version of the sentence with the complementizer that after the verb and one without it (64 sentences in total). The matrix subjects of the sentences were chosen to be minimally informative two-word noun phrases (e.g. *the men*, *they all*), to avoid biasing the

[3]We only considered active frames; after the verb has been read, passive frames such as *was discovered by* are no longer compatible with the sentence. Additionally, Gahl and colleagues distinguish frames that include participles (*Lola looked up from her book for the intransitive frame*) from frames that do not (*we watched attentively*); we ignored this distinction for the purposes of calculating subcategorization frame entropy.

Condition	Subcat. entropy	SC-surprisal	Frequency (log)
Low Entropy / Low Surprisal	1.13 (0.08)	1.52 (0.55)	3.64 (1.56)
Low Entropy / High Surprisal	1.09 (0.18)	4.15 (1.03)	4.31 (2.01)
High Entropy / Low Surprisal	1.7 (0.12)	1.58 (0.45)	3.6 (1.24)
High Entropy / High Surprisal	1.68 (0.17)	3.86 (0.79)	3.85 (1.25)

Table 3.1: Mean subcategorization entropy, SC surprisal and log-transformed frequency of the main verb in each of the conditions of the factorial design. We use Entropy to refer to high vs. low subcategorization frame entropy (i.e., single-step entropy at the verb) and Surprisal to refer to high vs. low sentential complement surprisal. Standard errors are shown in parentheses.

distribution over verb complement frames ahead of the verb. The same eight matrix subjects were used in all four conditions.[4] Following the complementizer (or the verb, if the complementizer was omitted) was a definite noun phrase (*the island*), which was always a plausible direct object of the verb (following Garnsey et al., 1997).[5] The frequency of this noun was matched across conditions.

The disambiguating region consisted of three words: either two auxiliary verbs (*had been*) or an auxiliary verb and negation (*might not*), followed by the past participle of a verb (*invaded*). Each of the function words appeared the same number of times in each condition. The verbs (*invaded*) were matched across conditions for frequency and length. The disambiguating region was followed by three more words, which were not analyzed.

[4]This meant that sentence subjects were repeated across items (4 times each). This choice was made in order to avoid more informative subjects. It is, however, theoretically possible that participants implicitly learned over the course of the experiment that these eight subjects were always predictive of an SC (Fine, Jaeger, Farmer, & Qian, 2013). We investigated whether the effects of surprisal and entropy change over the course of the experiment. We added list position to the region-byregion linear mixed-effects models. There were robust main effects of list position, such that participants became faster overall in later trials across regions ($ps < 0.001$). Crucially, however, the order effect did not interact with SC surprisal (all $ps > 0.1$). We conclude that there is no evidence that participants adapted their expectations over the course of the experiment.

[5]Most of the verbs were ambiguous between past tense and passive participle interpretations. In principle, this allows a reduced relative continuation; however, this continuation is very rare ($< 1\%$, cf. Fine et al., 2013), and unavailable for most of the verbs (e.g., *the men discovered the island [had been invaded]* cannot be interpreted as having the same structure as *the men sent the letter [were arrested]*).

In addition to the target sentences, the experiment included 64 filler sentences. These sentences contained various complex syntactic structures. The target sentences were separated from each other by at least one filler item. The first four trials always consisted of filler items to familiarize the participants with the task.

Eight experimental lists were created as follows. The 32 items were randomized such that sets of four consecutive items had one item of each condition (with fillers interspersed). The complementizer was omitted in every other item, counterbalanced across Lists 1 and 2. Lists 3 and 4 were obtained by reversing the order of presentation in Lists 1 and 2. The randomization procedure was then repeated to generate Lists 5 through 8. Each list was assigned to 16 participants.

3.2.2.3 Procedure

Sentences were presented word by word in a self-paced moving window paradigm (Just, Carpenter, & Woolley, 1982). After each trial, the participants were presented with a Y/N comprehension question to ensure that they were paying attention to the meaning of the sentence. Participants did not receive feedback on their responses. The experiment was conducted online using a Flash application written by Hal Tily. Participants took 17 minutes on average to complete the experiment (standard deviation: 4.1 minutes).

3.2.2.4 Preprocessing

Following standard procedure, individual words were excluded if their raw reading times (RT) were less than 100 ms or more than 2000 ms. All RTs were log-transformed to reduce right skew (Baayen & Milin, 2010; Frank, 2013). If a word's log-transformed RT was more than 3 standard deviations higher or lower than the participant's mean log-transformed RT, the word was excluded. RTs were then length-corrected by taking the residuals of a mixed-effects model which had log-transformed RT as the response variable, word length as a fixed effect, and a by-subject intercept and slope (following, e.g., Fine et al., 2013). Again following standard procedure, all trials including fillers were entered into the length correction model.

Two subjects were excluded because their answer accuracy was lower than 75%. The results reported in what follows are based on the remaining 126 subjects.

3.2.2.5 Statistical analysis

The resulting by-region length-corrected RTs were analyzed using linear mixed-effects models in R (Bates, Maechler, & Bolker, 2014), with crossed random effects for subjects and items. We used a maximal random effect structure: for items, a slope for sentence ambiguity; for subjects, slopes for all of the predictors and their interactions. In case the model fitting procedure did not converge, we removed the random slopes for the highest order interactions and refitted the model.[6] P-values for fixed effects were calculated using model comparison with a simpler model with the same random effect structure but without the fixed effect in question (following Barr, 2013).[7]

3.2.3 Results

3.2.3.1 Accuracy

Comprehension accuracy, including on fillers, was high (mean = 95.8%, standard deviation: 5.6%). To test whether accuracy differed between conditions, a mixed-effects logistic regression model (Jaeger, 2008) was fitted to the responses to the comprehension questions (excluding fillers). There were no significant main effects or interactions (all ps > 0.1, Wald statistic), indicating that ac-

[6] Due to model convergence issues, we had to exclude the random by-subject slopes for some of the interactions. Specifically, we excluded the slope for the three-way interaction in all four regions. For the matrix subject and disambiguating region we additionally excluded the bysubject slope for the SC-surprisal $\times$ Ambiguity interaction.

[7] Since the same set of eight NP subjects was used repeatedly in each of the four conditions, the NP subject can be seen as a random effect drawn from a population, and should arguably be modeled as such. We used forward model selection to test whether this random effect was necessary: for each region, we compared a model that included random intercepts and slopes for items and participants only (i.e. the models reported in the text) to a model that additionally included a random intercept and slopes for each predictor in our $2\times2\times2$ design. In all three regions (verb, ambiguous and disambiguating), adding this random effect did not improve the likelihood of the model significantly (all ps > 0.4).

curacy was similarly high across conditions. For the RT analyses, we analyzed all critical trials, regardless of accuracy.

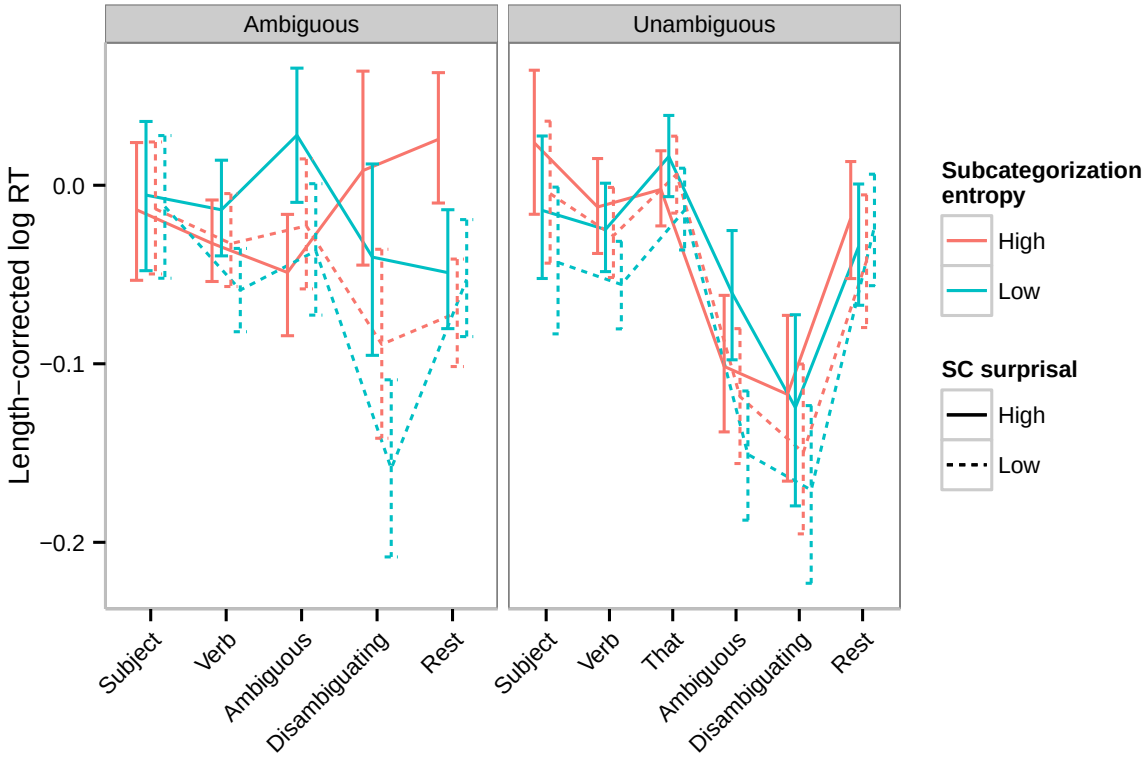


Figure 3.4: Mean reading times (RTs), averaged within regions. Error bars show bootstrapped 95% confidence intervals.

	Subject	Verb	Ambiguous	Disambiguating
Entropy	0.6	0.815	-0.95	1.03
Surprisal	0.87	1.61 (.)	1.58 (.)	2.17 (*)
Ambiguity	-0.02	-0.55	5.71 (***)	3.84 (***)
Entropy × Surprisal	-0.07	-1.22	-1.94 (*)	-0.32
Entropy × Ambiguity	-1.73 (.)	-1.07	-1.00	1.00
Surprisal × Ambiguity	-1.01	-0.02	-1.11	1.87 (.)
Entropy × Ambiguity × Surprisal	0.2	-0.8	-0.18	-0.14

Table 3.2: The table shows *t* statistics from a linear mixed-effects regression model. Entropy refers to high vs. low subcategorization frame entropy (i.e., single-step entropy at the verb); Surprisal refers to high vs. low sentential complement surprisal; and Ambiguity is positive for ambiguous sentences and negative for unambiguous ones. Legend: (***) $p < 0.001$, (**) $p < 0.01$, (*) $p < 0.05$, (.) $p < 0.1$.

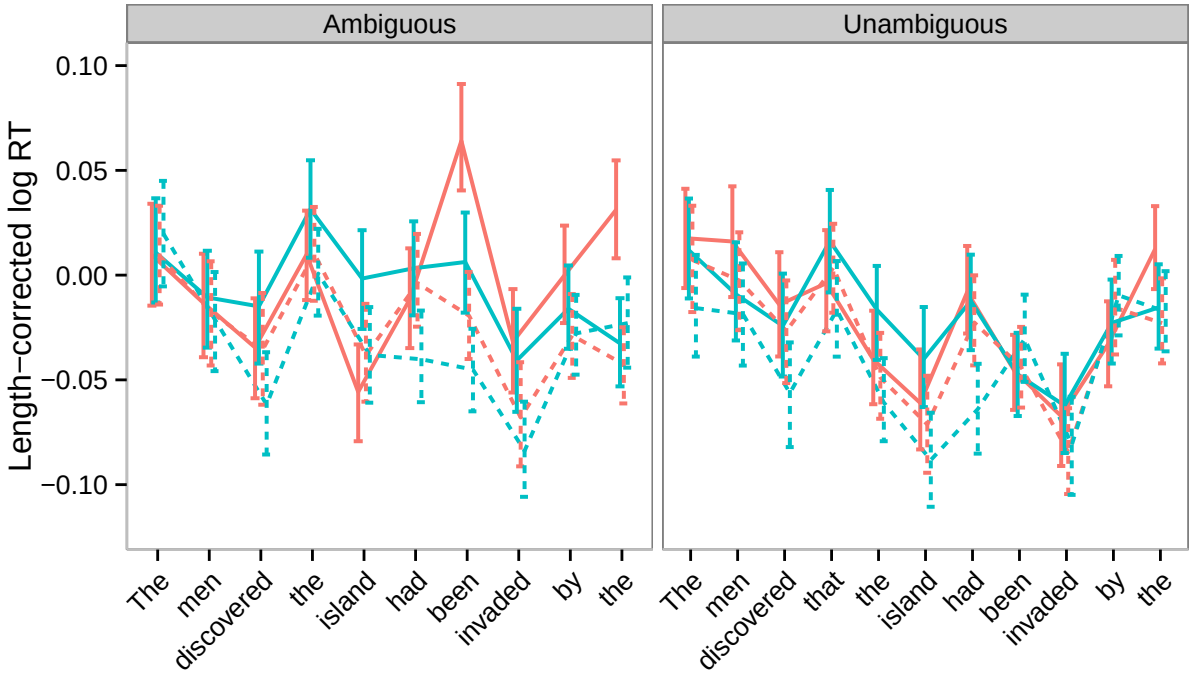


Figure 3.5: Mean reading times (RTs), word by word. Error bars show bootstrapped 95% confidence intervals.

3.2.3.2 Reading times

Mean RTs averaged within each region are shown in Fig. 3.4 (see also word-by-word RTs in Fig. 3.5). Following previous work (Garnsey et al., 1997), we split the sentences into four regions: subject, verb, ambiguous region and disambiguating region (see (2) above). Length-corrected RTs were averaged for each region of a given trial, and linear mixed effects models were fitted within each region. The results for all regions are summarized in Table 2.

Subject region: No effects reached significance (all ps > 0.1)

Verb region: Subcategorization entropy did not have a significant effect on RTs ($p > 0.5$). RTs in this region therefore do not support either the competition hypothesis or the entropy reduction hypothesis in their single-step form. Additional follow-up analyses that included a oneword spillover region in the verb analysis likewise failed to find an effect of entropy. Unexpectedly, low SC surprisal verbs were read more slowly, though this difference was only marginally significant ($p = 0.09$). None of the theories considered here predict an effect of SC surprisal at the verb: At the verb, readers do not yet know the category of the upcoming syntactic complement. We return to this trend when we discuss the full entropy analysis below, as it offers an alternative explanation for this apparent effect of SC surprisal.

Ambiguous region: The subject of the embedded clause (the "ambiguous region") was read faster in unambiguous sentences than in ambiguous sentences ($p < 0.001$). There was a marginal main effect of SC surprisal ($p = 0.1$). Simple effect analyses showed that RTs were significantly higher for high SC surprisal verbs than for low SC surprisal verbs in unambiguous sentences ($p = 0.04$) but not in ambiguous ones ($p > 0.4$). The interaction was not significant, however ($p > 0.2$). The effect of surprisal in unambiguous sentences may reflect spillover from the complementizer in unambiguous sentences, which is unpredictable after high SC surprisal verbs.

The main effect of subcategorization entropy was not significant ($p > 0.3$). There was a significant interaction between entropy and surprisal ($p = 0.04$). Simple effect analyses, collapsing

across the two levels of Ambiguity, showed that high SC surprisal verbs were associated with higher RTs when their subcategorization entropy was low ($p = 0.02$) but not when it was high ($p > 0.7$). The three-way interaction between surprisal, entropy and Ambiguity did not approach significance ($p > 0.5$).

Disambiguating region: This region was read faster in unambiguous sentences ($p < 0.001$). There was a main effect of SC surprisal ($p = 0.03$), as well as a marginally significant interaction between SC surprisal and ambiguity in this region ($p = 0.06$), such that the simple effect of SC surprisal in unambiguous sentences was not significant ($p > 0.2$), but the simple effect in ambiguous sentences was ($p = 0.007$). This is the signature expectation violation effect observed in previous studies (Garnsey et al., 1997; Trueswell et al., 1993). The main effect of entropy did not reach significance, and neither did any of the interactions between entropy and other predictors (ps > 0.3).

3.2.4 Discussion

The experiment replicated the SC surprisal effect found in previous studies (Garnsey et al., 1997; Trueswell et al., 1993): In ambiguous sentences, disambiguation in favor of an SC was more costly when the surprisal of an SC given the verb was high. Subcategorization entropy did not significantly affect RTs at the verb (or in any other region of the sentence). The absence of an entropy effect does not support either the competition or the entropy reduction hypotheses. One important caveat to this conclusion is that our experiment was based on the verbs' subcategorization entropy, that is, on readers' uncertainty about the syntactic category of the verb. As indicated in the introduction, this quantity does not take into account the reader's full uncertainty about the parse. We now examine the consequences of replacing subcategorization entropy with full entropy about the syntactic structure of the sentence.

3.3 Full entropy analysis

In order to assess the effect of the full uncertainty that a comprehender might experience during incremental sentence understanding, we derived full entropy estimates from a probabilistic context free grammar (PCFG) based on the Penn Treebank. With the exception of the matrix verbs, whose lexically-specific subcategorization probabilities are of direct relevance to this study, the grammar was unlexicalized: Its rules only made reference to parts-of-speech (syntactic categories) rather than individual lexical items. In contrast with the substantial transformations applied to the grammar in some state-of-the-art parsers (Johnson, 1998; Petrov & Klein, 2007), the grammar we used was very close to an untransformed "vanilla" PCFG. We made this decision to keep the grammar reasonably small, since computation of full entropy estimates becomes difficult with larger grammars (Roark et al., 2009). Appendix B provides more detail about the grammar.

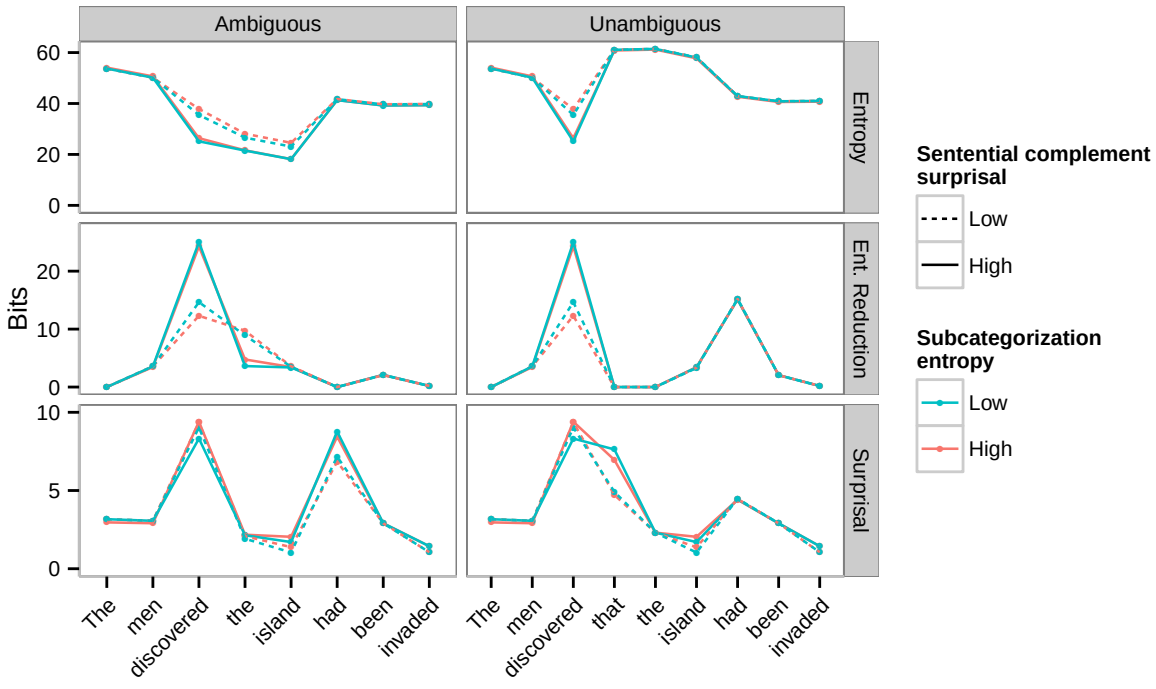


Figure 3.6: Word-by-word entropy, entropy reduction and surprisal predictions derived from the probabilistic context free grammar based on the Penn Treebank. Predictions are averaged within each of the four conditions of the factorial design. Following Hale (2006), we define entropy reduction at a word to be 0 when entropy after the word is greater than entropy before it.

The full entropy estimates derived from the grammar take into account not only the uncertainty about the next syntactic node, but also the uncertainty about the internal structure of that node (cf. Fig. 3.3). We use these full entropy estimates to test the competition and entropy reduction hypotheses, while simultaneously controlling for surprisal (estimated from the same PCFG). Fig. 3.6 summarizes the full entropy, entropy reduction and surprisal estimates derived from the PCFG, averaged within each of the four conditions of the factorial design (high vs. low SC surprisal $\times$ high vs. low subcategorization entropy). We go through each region separately. For each region, we begin by describing the relation between the factorial design and the PCFG-derived estimates. We then describe the predictions of the competition and entropy reduction hypotheses for RTs, based on the PCFG-derived estimates (see Table 3 for an overview). Finally, we analyze the effects of entropy and entropy reduction on RTs. Since entropy and entropy reduction tend to be highly correlated, we assess the effects of each of the variables in a separate model. Instead of the factorial SC surprisal and subcategorization entropy predictors, the models included continuous PCFG-based surprisal and one of the full entropy measures, as well as an interaction term. When not mentioned otherwise, all analyses contained the full random effect structure. All predictors were averaged across all of the words in a region, then centered and standardized before being entered into the model. Unless mentioned otherwise, collinearity did not play a role (correlations between predictors rs $< .5$).

3.3.1 Subject region

The same eight matrix subjects were used in all four conditions of the factorial design (see Fig. 3.6); the grammar-based predictors therefore did not differ across conditions in this region. The words that made up those eight subjects varied in their parts-of-speech: two people is a numeral followed by a plural noun (CD NNS in Penn Treebank notation), whereas the man is a determiner followed by a singular noun (DT NN). This resulted in some limited variability within each condition in the grammar-derived predictors for this region (Fig. 3.8). There was a strong correlation between entropy and entropy reduction ($r = -0.91$).

	Verb	Ambiguous	Disambiguating
Competition	Lower SC surprisal, higher subcat. entropy → higher full entropy → Lower full entropy reduction → shorter RTs ×	Longer RTs in unambiguous than in ambiguous sentences ×	
Entropy Reduction	Lower SC surprisal, higher subcat entropy → higher full entropy → Lower full entropy reduction → shorter RTs ✓	Shorter RTs in unambiguous than in ambiguous sentences ✓	Longer RTs in unambiguous than ambiguous sentences ∅
Within ambiguous sentences only:			
Competition		Lower SC surprisal, higher subcat entropy → longer RTs ∅	
Entropy reduction		Lower SC surprisal, higher subcat entropy → longer RTs ∅	

Table 3.3: Predictions made by the competition and entropy reduction hypotheses for the three main regions of the materials (the hypotheses do not make any predictions for the subject region). Predictions are shown both for the whole data set, focusing on the comparison between ambiguous and unambiguous sentences, and specifically for ambiguous sentences. Cells are empty whenever a hypothesis does not predict any RT difference in a region. Predictions confirmed by the results are marked with ✓; predictions not found confirmed are marked with ∅; predictions rejected by the results are marked with ×.

3.3.1.1 Results

Linear mixed-effects models did not yield any significant effects in this region, in either the entropy or the entropy reduction analysis (all ps > 0.3).

3.3.2 Verb region

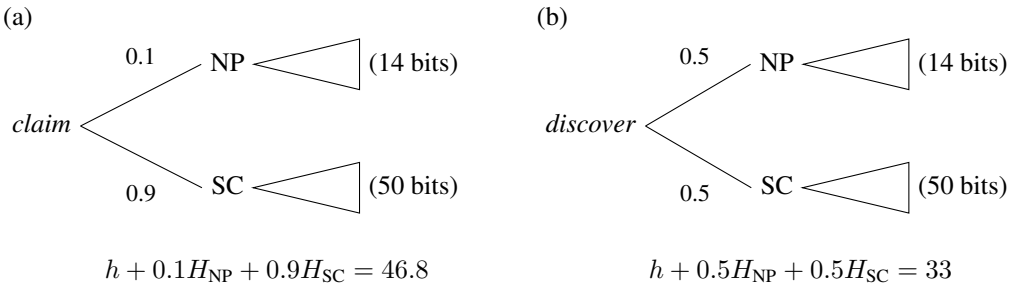


Figure 3.7: Illustration of the effect of the internal entropy of the verb's complements on full entropy after the verb. The internal entropy of an SC is much higher than both verbs' subcategorization entropy and the internal entropy of an NP; the most important predictor of full entropy in this case is therefore the probability of an SC (the specific values of internal entropy and subcategorization probabilities in the figure are for illustration purposes only).

Full entropy in this region is somewhat higher for verbs with high subcategorization entropy (Fig. 3.6). However, this difference is dwarfed by the substantial correlation between SC surprisal and full entropy: Verbs that are more likely to be followed by an SC (i.e., verbs with lower SC surprisal) have higher full entropy. This correlation, which may be unexpected at first blush, stems from the fact that full entropy at a given point in the derivation is calculated as the sum of single-step entropy and the expected full entropy of the structures that can be derived at that point (see Appendix B for details). SCs have many more potential internal structures than NPs or preposition phrases, and therefore higher internal entropy; when the probability of an SC is high, full entropy is dominated by the internal entropy of an SC (Fig. 3.7).

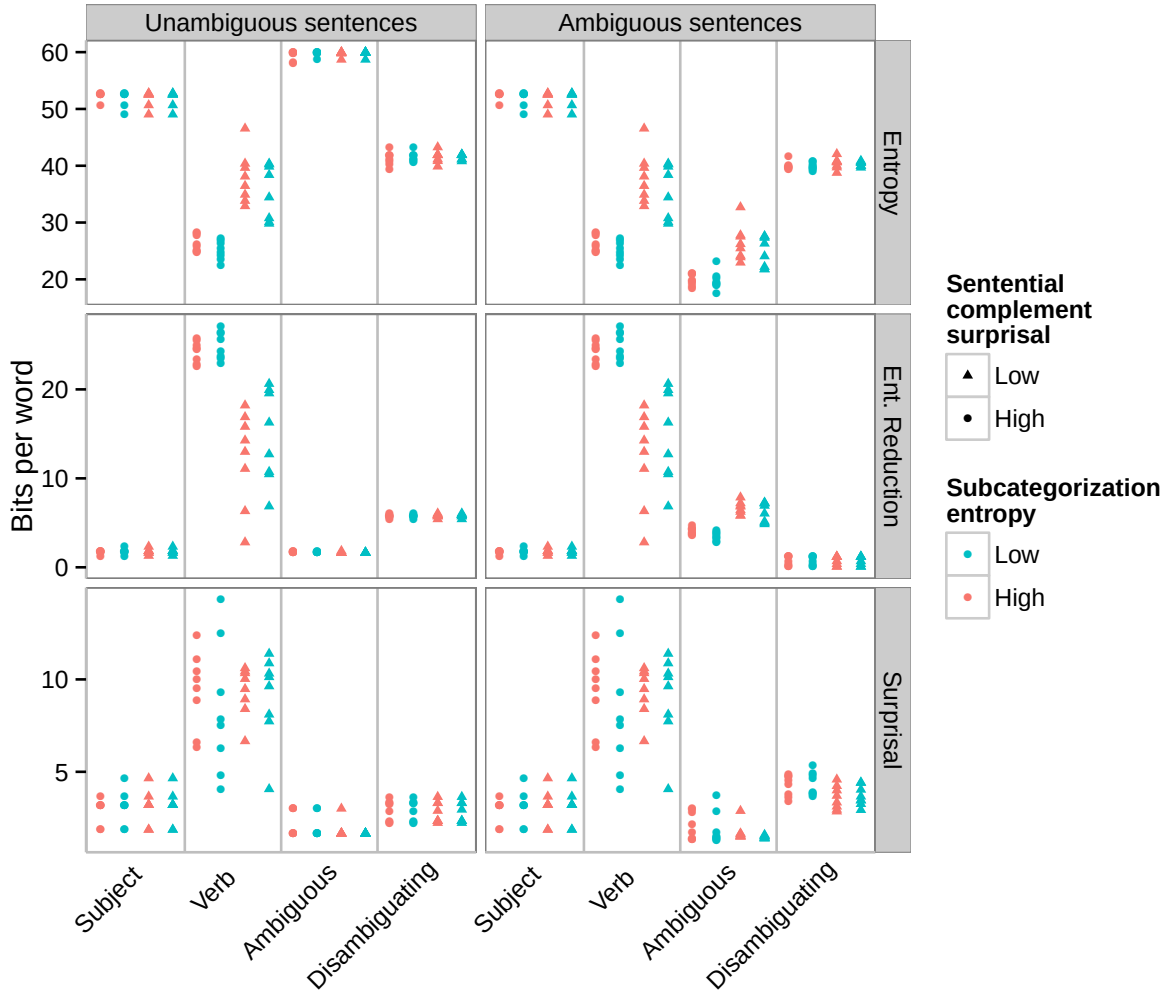


Figure 3.8: Variability across items of PCFG-derived variables (entropy, entropy reduction and surprisal). The values plotted are the mean of each variable across the region. Following Hale (2006), entropy reduction at a word is defined to be 0 when entropy after that word is greater than entropy before it.

Entropy and entropy reduction are highly inversely correlated in the verb region ($r = -0.98$), since entropy before the verb is similar across items (Fig. 3.8). As is the case in general, the competition hypothesis predicts a positive correlation between RTs entropy, and the entropy reduction hypothesis predicts a positive correlation with entropy reduction.

3.3.2.1 Results

The entropy reduction analysis found a positive effect of entropy reduction on RTs at the verb ($\beta = 0.014$, $p = 0.047$). The entropy analysis found a marginal negative effect of entropy ($\beta = -0.012$, $p = 0.08$). Neither analysis revealed any other effects (ps > 0.5). The effect of entropy reduction is in the direction predicted by the entropy reduction hypothesis; the direction of the entropy effect is opposite to that predicted by the competition hypothesis. To the extent that RTs on the verb region support either of the two hypotheses, they thus argue in favor of the entropy reduction hypothesis.

RTs on the verb region also shed light on the unexpected numerical difference in RTs between verbs with low and high SC surprisal verbs in the factorial (i.e., single-step entropy) analysis. SC surprisal is one of the major factors that determine full entropy at the verb. This suggests that the apparent effect of SC surprisal in the factorial analysis may be an artifact of its correlation with entropy reduction.

3.3.3 Ambiguous region

The high internal entropy of SCs continues to play a major role in the full entropy profile of the following regions as well. In unambiguous sentences, the sequence *that* + determiner increases the probability of an SC to 1, causing full entropy to rise sharply. In ambiguous sentences, on the other hand, some of the probability mass is reserved to the relatively low-entropy NP complement; this results in overall lower entropy. Counterintuitively, then, a strong version of the competition hypothesis, according to which predicted parses lead to competition prior to the presence of bottom-up evidence for those parses, would predict higher RTs in unambiguous sentences than in ambiguous sentences in this region. The entropy reduction hypothesis, on the other hand, predicts higher RTs in ambiguous than unambiguous sentences: Entropy decreases in ambiguous sentences, leading to some processing cost, whereas it increases in unambiguous sentences (recall that an increase in entropy is not predicted under the entropy reduction hypothesis to affect processing time).

Within ambiguous sentences, the predictions of the competition hypothesis for the ambiguous region are qualitatively similar to its predictions at the verb: The more likely the verb is to be followed by an SC, the higher the full entropy and hence the higher the RTs predicted by the competition hypothesis. The entropy reduction hypothesis predicts that this region should show the mirror image of the effect predicted at the verb: Verbs that strongly predict an SC cause a milder entropy reduction at the verb, but a steeper entropy reduction at the word the, which increases the probability of an NP and partly counteracts the verb-specific SC bias. At the ambiguous region of ambiguous sentences, then, both competition and entropy reduction predict the same qualitative effect across items: Lower SC surprisal should lead to increased RTs.

Within unambiguous sentences, neither theory predicts RT differences associated with the factorial conditions: Since the sentence has already been disambiguated in favor of the SC parse, entropy estimates for both this region and the word preceding it (*that*) do not depend on the matrix verb's subcategorization probabilities. The small differences that do exist across individual items (Fig. 3.8) stem from part-of-speech differences in the "ambiguous" noun phrase (plural vs. singular noun).

3.3.3.1 Results

Entropy and entropy reduction were highly correlated in the ambiguous region, though not quite as highly as in the verb regions ($r = -0.78$). We first conducted the analyses collapsing across ambiguous and unambiguous sentences. Entropy reduction correlated with increased reading times ($\beta = 0.042$, $p < 0.001$), and entropy correlated with decreased reading times ($\beta = -0.043$, $p < 0.001$). No other effects reached significance in either analysis ($ps > 0.5$). These results are predicted by the entropy reduction hypothesis, and opposite to the predictions of the competition hypothesis. Since entropy effects in the ambiguous region are highly confounded with Ambiguity (entropy: $r = 0.99$; entropy reduction: $r = -0.85$), however, this result has to be interpreted with caution. For example, it is possible that the observed effect stems from other differences between the ambiguous and unambiguous sentences, such as the presence of temporary ambiguity.

To test whether the effect of the entropy measures was solely carried by Ambiguity, we repeated both analyses over ambiguous sentences only (there was little variability across items in unambiguous sentences; see Fig. 3.8). None of the predictors in either the entropy reduction model or the entropy model had a significant effect on RTs (all ps > 0.3). This suggests that the effects found in the first analyses (over ambiguous and unambiguous sentences) were driven by difference between ambiguous and unambiguous sentences. It leaves open, however, whether the lack of any effects within ambiguous sentences is due to the reduced variability in entropy and entropy reduction within ambiguous, as compared to between ambiguous and unambiguous sentences (cf. Fig. 3.8).

3.3.4 Disambiguating region

When readers reach the disambiguating region, they have ample evidence that the high entropy SC parse is the correct one. This leads to high overall entropy, regardless of whether the sentence was originally ambiguous or unambiguous and of expectations derived from the verb's subcategorization bias. Consequently, the competition hypothesis predicts no systematic difference across sentence types at this region.

Conversely, the entropy reduction hypothesis predicts that the disambiguating region should be read faster in ambiguous compared to unambiguous sentences. This somewhat counterintuitive prediction deserves some elaboration. When ambiguous sentences are disambiguated in favor of the SC parse, entropy increases sharply because of the higher internal entropy of SCs. The degree to which it increases depends on the verb's subcategorization bias; however, the entropy reduction hypothesis predicts processing difficulty only for decreases in entropy, and hence never predicts any processing cost at the disambiguation region of ambiguous sentences. For unambiguous sentences, on the other hand, readers already know that the complement is an SC. This means that entropy will go down at the first word of the disambiguating region, because on average additional words reduce entropy (see also Fig. 3.6). This reduction in entropy, however mild, entails some processing cost, compared to no processing cost at all in ambiguous sentences.

Neither hypothesis predicts any difference based on the matrix verb's subcategorization distribution, even within ambiguous sentences: While entropy before the disambiguation point does differ across items, it always increases at the first word of the disambiguating region. The entropy reduction hypothesis therefore does not predict any RT difference across items at this word. The differences in entropy before the disambiguation point do not affect the predictions of the competition hypothesis either, since this hypothesis does not take history into account. From the disambiguation point on, all items have an SC parse. Consequently, word-by-word entropy and entropy reduction estimates will only vary with the syntactic categories of the verbal complex in the disambiguating region, e.g., *had been invaded* (VBD VBN VBN) vs. *should be reported* (MD VB VBD).

Finally, surprisal predicts that the disambiguating region should be read more slowly in ambiguous sentences, where it is in fact disambiguating.

3.3.4.1 Results

Following our analyses of the ambiguous region, we excluded the Ambiguity factor from the analysis (i.e., collapsed all sentences). Surprisal had a significant effect in both the entropy analysis ($\beta = 0.06$, $p = 0.01$) and the entropy reduction analysis ($\beta = 0.05$, $p = 0.03$). None of the other predictors were significant (ps > 0.1).

3.3.5 Summary of full entropy analyses

A summary of the predictions of the competition and entropy reduction hypotheses and the results of the full entropy analyses is given in Table 3. We found a significant effect of entropy reduction on RTs at the verb: Higher entropy reduction at the verb correlated with longer RTs. The interpretation of this effect is somewhat complicated by the fact that we also observed a marginal effect of entropy in the same region. The high correlation between entropy and entropy reduction makes it impossible to distinguish these two effects via model comparison. However, the direction of the observed pattern is consistent with the predictions of the entropy reduction hypothesis and

inconsistent with the predictions of the competition hypothesis. Similarly, RTs at the ambiguous region support the predictions of the entropy reduction hypothesis but are inconsistent with the predictions of the competition hypothesis. RTs on the verb and ambiguous region thus provide support for the entropy reduction hypothesis over the competition hypothesis, while comparing both hypotheses on equal ground (using the same representational assumptions, the same stimuli, and the same control factors).

It is important to keep in mind that syntactic structure is kept constant only at the verb; when both Ambiguity conditions are included in the analysis of the ambiguous region, the regression model collapses across different syntactic structures (much like the comparison between subject and object relative clauses in Yun et al., 2015). Consequently, the longer reading times in the ambiguous region of ambiguous sentences may reflect an unrelated factor, such as the fact that the ambiguous region follows a frequent function word in unambiguous sentences (*that*), but an infrequent content word in ambiguous sentences (namely, the verb; cf. Clifton & Staub, 2008). The results at the verb therefore constitute stronger support for the role of uncertainty than the results at the ambiguous region.

RTs at the disambiguating region were predicted by neither the competition nor the entropy reduction hypothesis; the only significant predictor of reading times in this region was surprisal. This result has to be interpreted with caution, however: Surprisal was highly correlated with both entropy ($r = -0.79$) and entropy reduction ($r = -0.71$). This means that surprisal and entropy reduction effects are predicted to operate in the opposite direction from each other. It is thus possible that the strong effects of surprisal masks any effect of entropy reduction.

3.4 General discussion

Building on recent evidence that readers maintain expectations over upcoming syntactic structure, this study has investigated how readers' parsing performance is affected by the probability distribution of those expectations, focusing specifically on uncertainty about upcoming structure. We

outlined two hypotheses about the potential role of uncertainty in parsing: the competition hypothesis, according to which higher uncertainty should result in the activation of multiple structures that compete with each other, thereby slowing down processing (Elman et al., 2005; McRae et al., 1998); and the entropy reduction hypothesis, according to which processing is slowed down by any word that reduces uncertainty (J. Hale, 2006).

We assessed uncertainty about the parse in two ways: single-step entropy, which quantifies uncertainty about the next derivation step, in this case the category of the verb's complement (subcategorization frame); and full entropy, which quantifies uncertainty about the syntactic structure of the whole sentence. Much previous work has employed single-step or other non-full estimates of entropy (e.g., Frank, 2013; Roark et al., 2009; Wu et al., 2010; but see J. Hale, 2003b, 2006). This is potentially problematic as the entropy reduction hypothesis is formulated in full entropy terms (J. Hale, 2006).

Indeed, the distinction between single-step and full entropy turned out to be critical for the current study: Single-step entropy did not affect RTs (aside for an unexpected and likely spurious interaction with surprisal in the ambiguous region), but full entropy did. RTs were longer when post-verb full entropy was lower. The direction of the effect is not compatible with our implementation of the competition hypothesis, according to which higher entropy should lead to increased competition and slower processing. It is, however, consistent with the entropy reduction hypothesis: Entropy before the verb was always higher than entropy after it; if postverb (full) entropy is high, then, the verb did not reduce entropy by much, and thus is (correctly) predicted to be relatively easy to process. The entropy reduction hypothesis also correctly predicts that the ambiguous region of unambiguous sentences should be read faster, compared to ambiguous sentences.

The only prediction of the entropy reduction hypothesis that was not confirmed applies to the disambiguation region. Here, the entropy reduction hypothesis predicts that ambiguous sentences should be read faster than unambiguous ones. The opposite was observed. As outlined above, however, this does not necessarily provide a strong argument against the entropy reduction hypothesis: In the disambiguation region, surprisal is expected to have the opposite effect from entropy reduc-

tion. Given the strong surprisal effects at the disambiguating region, all that can be concluded from this result is that (in this case) surprisal may have a stronger effect on processing difficulty than entropy reduction. Indeed, while both surprisal and entropy reduction had statistically significant effects on reading times – at the disambiguating region and at the verb, respectively – the size of the surprisal effect was larger than the size of the entropy reduction effect ($\beta = 0.06$ vs. $\beta = 0.014$).

In summary, it seems that (at least) both surprisal and the entropy reduction hypothesis are required to account for our results, whereas no support for the competition hypothesis (as formulated here) was observed.

In the remainder of the discussion, we review the role of lookahead in human language processing. The current study has evaluated the two ends of the spectrum: $n = 1$ (single-step entropy) and $n = \infty$ (full entropy). Other values of n are also possible (indeed, likely); our finding that RTs can be predicted by full but not single-step entropy further supports the conclusion that human parsing during reading involves lookahead of at least several derivation steps. This conclusion is in line with the conclusions of Frank (2013) lookahead distances of 1 to 4 steps and found that increasing the amount of lookahead increased the extent to which entropy reduction predicted RTs.[8]

Changing the lookahead distance may qualitatively change the predictions made by the competition and entropy reduction hypotheses. Consider, for example, the predictions of the competition hypothesis (as implemented here) for the ambiguous region. At first blush, one might expect that ambiguity will lead to more competition because of the uncertainty about the category of the complement (cf. Green & Mitchell, 2006; Levy, 2008a, p. 1152). However, as we have outlined above, there are actually two components that combine to determine uncertainty (competition) at this point in the sentence: the uncertainty about the category of the complement (e.g., whether it is an SC) and the uncertainty about the internal structure of the complement.

[8]The lookahead distance in Frank (2013) is not directly comparable to ours. Frank calculated entropy based on word predictions derived from a connectionist network; a lookahead of $n = 4$ in that model corresponds to predicting the next four words. Conversely, our model predicts PCFG rewrite rules, not words; multiple PCFG derivation steps may be required to predict each word (and vice versa), such that four words can correspond to three, five or eight PCFG rewrite rules.

Under the infinite lookahead assumption, the latter turns out to dominate the former. The competition hypothesis therefore predicts that the disambiguating region will be processed more slowly in unambiguous sentences than in ambiguous ones – contrary to our findings (which replicate Kennison, 2001; Pickering & Branigan, 1998). However, since much of the large entropy associated with SCs comes from their internal structure, shorter lookahead distances would decrease the relative contribution of the internal structure of SCs to the overall uncertainty experienced in the ambiguous region, bringing the predictions of the competition hypothesis in line with the empirical findings. Determining the appropriate lookahead distance therefore constitutes an interesting question for future computational studies.

It is also important to point out that entropy estimates and the definition of what constitutes a single derivation step may depend on the strategy employed by the parser and on the precise representation of the grammar. For example, a parser may choose to defer the prediction of an NP or SC category until there is some information supporting either of these categories (in a top-down parser, this strategy could be implemented by applying a right-binarization transform, which underspecifies the category of the complement; Roark & Johnson, 1999). Such a parsing strategy may predict no uncertainty at all at the verb. Furthermore, the grammar representation we employed was based on the Penn Treebank (Marcus et al., 1993), with minimal modifications (see Appendix B). The Penn Treebank has a small nonterminal set (around 20 nonterminals). Larger nonterminal sets, created by splitting existing symbols into finer-grained categories (e.g., by annotating a node in the tree with the tags of its siblings and parents), have been shown to provide a more realistic probabilistic model of natural language syntax (Johnson, 1998; Klein & Manning, 2003; Petrov & Klein, 2007). Entropy estimates and lookahead distance based on narrower categories are likely to differ significantly from those based on broad categories: Single-step prediction of a narrow category can approximate several stages of prediction of broader categories. More generally, grammatical formalisms that allow for some degree of context-sensitivity (e.g., Kallmeyer, 2010) have been argued to be more adequate models of human language syntax. In future work, it is worth exploring

how these grammatical representational assumptions affect entropy estimates in general and the distinction between single-step and full entropy in particular.

3.5 Conclusion

This study used syntactic expectations induced by individual lexical items to examine the role of uncertainty over expectations in parsing. The results lend some support to the entropy reduction hypothesis (J. Hale, 2006). The design of the current study addressed differences between previous works that complicated an evaluation of the entropy reduction and competing hypotheses (cf. Levy & Gibson, 2013). However, the entropy reduction hypothesis failed to predict RTs where it was in conflict with the surprisal hypothesis (J. Hale, 2001; Levy, 2008a). This suggests that predictability (surprisal) and uncertainty both play a role in explaining processing difficulty in sentence processing. Modeling of the RT results further suggests that the extent to which uncertainty predicted processing difficulty depended on the depth of syntactic lookahead that readers were assumed to perform: Uncertainty was not a significant predictor of RTs when only the syntactic category of the verb's complement was considered, and became significant only when the internal complexity of the complement was taken into account.

3.6 Appendix A: List of materials

3.6.1 Low subcategorization entropy, low SC surprisal

1. The men discovered (that) the island had been invaded by the enemy.
2. The women revealed (that) the secret had been exposed by the officials.
3. The man noticed (that) the mistake had not happened due to negligence.
4. The woman assumed (that) the blame might have belonged to the driver.
5. They all indicated (that) the problem might not bother the entire team.

6. Two people found (that) the equipment should be reported stolen right away.

7. Some people sensed (that) the conflict should be resolved quickly and peacefully.

8. Many people guaranteed (that) the loan would be paid off on time.

3.6.2 Low subcategorization entropy, high SC surprisal

1. The woman determined (that) the estimate had been inflated by the accountant.

2. Two people heard (that) the album had been criticized in the magazine.

3. Some people understood (that) the message had not meant much to foreigners.

4. They all read (that) the newspaper might be going out of business.

5. The women worried (that) the parents might have become quite restless recently.

6. Many people advocated (that) the truth should be made public without delay.

7. The man taught (that) the children should be sheltered from all harm.

8. The men projected (that) the film would not gross enough in cinemas.

3.6.3 High subcategorization entropy, low SC surprisal

1. They all claimed (that) the luggage had been stolen from the hotel.

2. Some people regretted (that) the decision had been reached without any discussion.

3. The men remembered (that) the appointment had not changed since last week.

4. The women warned (that) the drivers might have drunk too much vodka.

5. Many people feared (that) the future might not hold hope for them.

6. The man proposed (that) the idea should be abandoned for financial reasons.

7. Two people suggested (that) the scene should be filmed right before sunset.

8. The woman announced (that) the wedding would be postponed until late August.

3.6.4 High subcategorization entropy, high SC surprisal

1. The men forgot (that) the details had been worked out in advance.

2. The man observed (that) the patient had been sent home too early.

3. The woman recalled (that) the speech had not gone over very well.

4. The women answered (that) the questions might be discussed during the meeting.

5. Some people added (that) the numbers might have decreased since last year.

6. Two people wrote (that) the interview should be conducted over the phone.

7. Many people advised (that) the president should be considering further budget cuts.

8. The men begged (that) the judge would not treat the defendant harshly.

3.7 Appendix B: Grammar definition and estimation

3.7.1 Definitions

A probabilistic context free grammar (PCFG) consists of:

- A set of non-terminals V, which includes intermediate categories such as VP (verb phrase) and N (noun);

- A set of terminal symbols T, which represent specific words (e.g., *dog*);

- A special start symbol S

- A set of rule productions of the form $X \rightarrow \alpha$, where X is a nonterminal and α is a sequence of terminals or nonterminals (e.g., VP $\rightarrow$ V NP)
- And a function ρ that assigns a probability to each production rule.

We define R(X) to be the set of rules rewriting nonterminal X.

A PCFG is considered lexicalized if some of its nonterminals include lexical annotations that allow the identity of a lexical head to affect the probability of modifiers that co-occur with it. For example, a lexicalized grammar can have both ρ(VP[break] $\rightarrow$ V[break] NP) $= 0.3$ and ρ(VP[hit] $\rightarrow$ V[hit] NP) $= 0.7$.

We first define the entropy of the next derivation step (single-step entropy). If $a_i \in V$ is a nonterminal (e.g., VP), the single-step entropy $h(a_i)$ corresponding to a_i is given by

$$h(a_i) = -\sum_{r \in R(a_i)} \rho(r) \log_2 \rho(r) \tag{3.2}$$

The full entropy of a nonterminal is defined recursively, as the sum of two terms: the single-step entropy of the nonterminal and the expected sum of the full entropy of any nonterminals that can be derived from the nonterminal. Formally, the full entropy $H(a_i)$ of nonterminal a_i is given by (Grenander, 1967):

$$H(a_i) = h(a_i) + \sum_{r \in R(a_i)} \rho(r) \sum_{j=1}^{k_r} H(a_{r,j}) \tag{3.3}$$

Where $a_{r_1}, \ldots, a_{r_{k_r}}$ are the nonterminals on the right hand side of rule r. The closed form formula for the recursion is:

$$H = (I - A)^{-1} h \tag{3.4}$$

Where $H = (H_1, \ldots, H_{|V|})$ is the vector of all full entropy values, $h = (h_1, \ldots, h_{|V|})$ is the vector of all single-step entropy values, I is the $|V| \times |V|$ identity matrix, and A is a $|V| \times |V|$

matrix, in which the element in row i and column j indicates the expected count of instances of nonterminal a_j resulting from rewriting nonterminal a_i.

3.7.2 Full entropy estimation

Word-by-word entropy estimates for our materials were derived using the Cornell Conditional Probability Calculator (Z. Chen, Hunter, Yun, & Hale, 2014) from a probabilistic context-free grammar estimated from the Penn Treebank. Following standard practice, we removed grammatical role and filler-gap annotations (e.g., NP-SUBJ-2 was replaced by NP). We reduced the size of the grammar by removing rules that included punctuation, rules that occurred less than 100 times (out of the total 1320490 nonterminal productions) and rules that had a probability of less than 0.01. These steps resulted in the removal of 13%, 14% and 10% rule production tokens, respectively.

The grammar was unlexicalized, except for verb-specific production rules that captured the differences in subcategorization probabilities among the 32 verbs in the experiment (again based on Gahl et al., 2004). Half of the probability mass from all (unlexicalized) rules deriving VP was divided among the lexicalized rules. The conditional probability of each rule was proportional to the verb's frequency. For example, the probability of the rule VP[discover] $\rightarrow$ V[discover] NP was defined to be:

$$\rho(\text{VP[discover]}) \rightarrow \text{V[discover] NP}) = \frac{1}{2}\frac{freq(\text{discover})P(\text{NP}|\text{discover})}{\sum_i freq(v_i)} \tag{3.5}$$

Unlexicalized grammars read off a treebank have been shown to make excessively strong assumptions of context-freeness, which affects the accuracy of probability estimates derived from such grammars (Johnson, 1998). The adequacy of the grammar is typically improved either by lexicalizing the grammar or by adding contextual information to some of the tags; e.g., the NP tag may be split into NP^VP for an NP whose parent is an VP and NP^S for an NP whose parent is an S (Johnson, 1998; Klein & Manning, 2003). At the same time, the number of nonterminals

in the grammar had to be kept to a minimum to enable the use of the closed form full entropy formula, which requires inverting a matrix that has as many rows as the number of nonterminals in the grammar. We therefore only split the tags that were most relevant to the probability estimates derived for the experimental materials.

First, the word that is tagged in the Penn Treebank as a preposition (IN) when it occurs as a subordinating conjunction (as in *the men discovered that*...). This resulted in SCs being erroneously parsed as prepositional phrases. We therefore replaced the generic tag IN with IN[that] in those cases; similarly, for auxiliary verbs we replaced VBD with VBD[had] and VBN with VBN[been].

Second, the grammar assigned implausibly high probabilities to reduced relative readings of the materials (where *discovered the island* is attached as a modifier of *the men*, by analogy with *the men discovered by the police*). Since the verb in a reduced relative must be a past participle (VBN), we split the VP category into subcategories based on the VP's leftmost child, e.g., VP_VBD is a VP headed by a past-tense verb (VBD), such that only a VP_VBN can modify a noun. We likewise split SBAR into SBAR[overt] when the SBAR had an overt complementizer and SBAR[none] when it did not.

The focus of this study is on syntactic surprisal and entropy, that is, on the portion of those measures that is due to the part-of-speech of the word rather than its identity (Roark et al., 2009). To tease out the syntactic component of these measures, then, the input to the parser consisted of parts-of-speech, with the exception of the matrix verb; e.g.:

(6)

The	men	discovered	the	island	had	been	invaded	by	the	enemy.
DT	NNS	discovered	DT	NN	VBD[had]	VBN[been]	VBN	IN	DT	NN

4

Syntactic context effects in visual word recognition: An MEG study

4.1 Introduction[1]

Research on isolated word recognition has uncovered an array of lexical, orthographic and semantic factors that affect the word recognition process (see Balota, Yap, & Cortese, 2006 for a review). Many of these factors are properties of the specific form being recognized, such as its frequency or its length. Some, however, are properties of the environments in which the recognized form is typically embedded. The role of context has been most thoroughly investigated in the case of word-internal structure (morphology). For instance, it has been shown that the speed of recognition of a monomorphemic word, such as *look*, is modulated by the number and frequency of words that are derived from it, in this case words like *looked* or *looking* (Schreuder & Baayen, 1997; Baayen et al., 2006).

Research in theoretical linguistics indicates that word-internal structure and word-external structure have much in common (Halle & Marantz, 1993, 1994). It is therefore natural to ask whether word-internal morphological context effects extend to word-external context: the words and structures that typically surround the recognized word in texts. This indeed turns out to be

[1]This work was done in collaboration with Alec Marantz and Liina Pylkkänen and was published in *The Mental Lexicon* in 2013.

case. S. A. McDonald and Shillcock (2001) found that words were responded to more slowly in isolation if they occurred in an unusual set of sentential contexts compared to the typical contexts in the language. More recently, contextual effects in lexical decision have been reported for the distribution of prepositions and adjectives preceding a noun: Nouns that co-occur with an unusual set of prepositions take longer to recognize (Baayen, Milin, Djurdjević, Hendrix, & Marelli, 2011).

Two types of context could potentially have an effect on word recognition: collocational context and syntactic context. We define the collocational context of a word as comprising the specific lexical items (*blue*, *dog*) that tend to co-occur with the recognized word, regardless of the syntactic structure of the sentence. By contrast, the syntactic context of the word abstracts away from particular lexical items, focusing instead on the syntactic representation of the phrases that the word appears in: Is it usually modified by an adjective? Does it tend to be followed by a verb?

Previous studies of contextual effects in word recognition either have not attempted to dissociate collocational and syntactic context (S. A. McDonald & Shillcock, 2001), or explicitly controlled for the syntactic environment in order to isolate the collocational context (Baayen, 2010; Baayen et al., 2011). Baayen et al. (2011), for example, limited their definition of context to the prepositions that preceded the noun being recognized, ignoring verbs, adjectives and other syntactic categories. It is therefore unknown whether syntactic context affects word recognition in isolation. Furthermore, contextual effects have only been reported for reaction times in behavioral experiments, and it is not known whether or how they modulate neural activity. This study fills both of these gaps, by measuring the effects of both syntactic and collocational context in a lexical decision task, while recording neural activity with magnetoencephalography (MEG).

To examine the effects of syntactic context, we exploit the fact that verbs vary in the types of syntactic phrases they can take as their complements (their *subcategorization frames*, Chomsky, 1965). The verb *devour*, for example, is always followed by a noun phrase, *dine* is never followed by a noun phrase, and *eat* can appear either with a noun phrase or without one:

(7) a. We ate.

b. We ate the turkey.

c. *We devoured.

d. We devoured the turkey.

e. We dined.

f. *We dined the turkey.

Verbs also differ in the statistical distribution of their subcategorization frames (SCF). Both *accept* and *prove*, for example, can occur with either a noun phrase (NP) or a subordinate clause (SC), yet they differ in the relative frequencies of these two frames (Garnsey et al., 1997):

(8) a. $P(\text{NP}|\textit{accept}) = 0.98$: He accepted the proposal.

b. $P(\text{SC}|\textit{accept}) = 0.01$: He accepted that he was wrong.

c. $P(\text{NP}|\textit{prove}) = 0.23$: He proved the claim.

d. $P(\text{SC}|\textit{prove}) = 0.61$: He proved that I was wrong.

Language comprehenders are aware of the distribution of verbs' subcategorization frames, and use this information to make predictions about the syntactic category of the verb's complement during sentence processing (Garnsey et al., 1997; M. Wilson & Garnsey, 2009; Arai & Keller, 2013).

Subcategorization distributions must be represented as vectors, with each component of the vector corresponding to the probability of a given subcategorization frame. It is not immediately obvious how to relate these vectors to dependent measures such as reaction times or neural activity. We explore two different ways to summarize a verb's SCF distribution as a single quantity: first, the entropy of the distribution, and second, its relative entropy compared to the overall distribution of SCFs in English. The entropy of the SCF distribution is a combined measure of the number of possible frames and the extent to which their distribution is balanced, reflecting the degree of uncertainty about the syntactic category of the verb's complement (Moscoso del Prado Martín et al., 2004). In the case of a verb that only allows one type of syntactic complement, there is no uncertainty at all as to the category of its complement, so the verb's SCF entropy is equal to 0.

Among verbs with two possible SCFs, entropy will be higher when the two are equally likely. Conversely, when one of the frames is much more likely than the others, the entropy will be close to 0. Finally, a verb with three equally distributed frames will have higher entropy than a verb with only two equally distributed frames. Mathematically, if a verb X has n possible frames, and the probability of the i-th frame is p_i, its SCF entropy will be:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{4.1}$$

Higher entropy has been associated with shorter lexical decision reaction times in the morphological domain. A word's morphological family is defined as the set of complex words in which the word appears as a constituent (Schreuder & Baayen, 1997). This set can be divided into an inflectional family (*thinks*, *thinking* for the base word *think*), and a derivational family (*thinker*, *rethink*). Both derivational family entropy and inflectional family entropy have a facilitatory effect on visual lexical decision (Baayen et al., 2006). Likewise, high entropy over a word's morphological continuations facilitates reaction times in auditory lexical decision (Baayen, Wurm, & Aycock, 2007; Wurm et al., 2006).

In other domains, higher entropy may result in a processing slowdown. Studies of lexical ambiguity, for example, have shown that words that have two unrelated meanings take longer to respond to in a lexical decision task than unambiguous words (Rodd, Gaskell, & Marslen-Wilson, 2002; Beretta, Fiorentino, & Poeppel, 2005). Moreover, ambiguous words take longer to process when both of the meanings of the word are equally frequent (high meaning entropy) than when one of the meanings is dominant (low meaning entropy) (Duffy et al., 1988). These results have been attributed to competition between the two meanings of the ambiguous word: since the meanings inhibit each other, the semantic activation associated with the word does not reach the threshold required to make a lexical decision. Similar results have been reported in fMRI, where ambiguous words elicited increased BOLD signal in language areas (Rodd, Davis, & Johnsrude, 2005). Finally, in MEG, ambiguous words with high meaning entropy lead to increased neural activity compared to low meaning entropy words (Simon et al., 2012). This suggests that

competition leads to increased MEG signal, though in some studies competition modulated the latency rather than the amplitude of the neural response (Beretta et al., 2005; Pylkkänen et al., 2004).

In summary, higher SCF entropy may have either a facilitatory or inhibitory effect: facilitatory if subcategorization frames behave like morphological continuations, and inhibitory if they behave like meanings competing for selection.

SCF entropy is a property of the SCF distribution of a single verb. By contrast, relative SCF entropy quantifies the divergence between the verb's specific distribution and the distribution of the "average" English verb, obtained by collapsing together the SCF distributions of all verbs in the language. Assume, for example, that English only had three subcategorization frames: prepositional phrase (*talk about something*), noun phrase (*break something*) and the intransitive frame (*snore*), and that their overall probabilities in the language were 0.2, 0.4 and 0.4, respectively. If *break* has the SCF distribution (0.2, 0.45, 0.35), which is similar to the overall distribution, then its relative SCF entropy will be low. On the other hand, if *snore*, which is only compatible with the intransitive frame, has the distribution (0, 0, 1), then its relative entropy will be fairly high.

Formally, if the overall probability of frame i in the language is q_i, and its probability given the verb X is p_i, then verb's SCF relative entropy is given by:

$$\sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \tag{4.2}$$

In the morphological domain, nouns with high relative inflectional entropy are responded to more slowly in lexical decision (Milin, Filipović Djurdjević, & Moscoso del Prado Martín, 2009). High divergence between the collocational context of the word being recognized and the average collocational context has a similar inhibitory effect (S. A. McDonald & Shillcock, 2001; Baayen et al., 2011). The prediction is therefore that higher SCF relative entropy will result in higher processing load.

The experiment described in this paper investigated the effect of the two subcategorization variables on participants' neural activity while they were performing lexical decision task on a set

of verbs. Source localization techniques were used to determine which brain regions generated the observed MEG signal at each timepoint (Hämäläinen, Hari, Ilmoniemi, Knuutila, & Lounasmaa, 1993). We used a region of interest analysis to reduce the dimensionality of the data and incorporate the results of prior research. Our primary region of interest was the left lateral anterior temporal lobe (ATL), which we define as the parts of the superior temporal gyrus (STG) and middle temporal gyrus (MTG) that lie anterior to the auditory cortex (see Figure 4.1). ATL lesions are associated with impaired performance on basic morphosyntactic tasks (Dronkers, Wilkins, Van Valin, et al., 2004). This region shows greater neural activity on two-word phrases compared to individual words (Bemis & Pylkkänen, 2011), and is consistently more active when processing sentences than when processing unstructured word lists, across techniques and modalities (Brennan & Pylkkänen, 2012; Humphries, Binder, Medler, & Liebenthal, 2006; Mazoyer et al., 1993). In MEG studies, combinatory effects in the ATL typically appear between 200 ms and 300 ms (Bemis & Pylkkänen, 2011, 2013; Brennan & Pylkkänen, 2012). Finally, multi-voxel pattern analysis of fMRI data has shown differential ATL activity patterns corresponding to different argument structure realizations of the same verb (Allen, Pereira, Botvinick, & Goldberg, 2012). The ATL therefore emerges as the region most likely to be sensitive to the properties of the immediate context of a word, and specifically to subcategorization frames.

We also report results from two other left-hemisphere language regions: the posterior temporal lobe (PTL), which includes the parts of the STG and MTG posterior to the auditory cortex; and the posterior inferior frontal gyrus (Broca's area), which includes Brodmann areas 44 and 45. PTL lesions are associated with word-level impairments, and damage to Broca's area is associated with impaired processing of syntactically complex sentences (Dronkers et al., 2004). The functional dissociation of the three regions is not clear-cut, however. In addition to its involvement in composition of words into phrases, the ATL is recruited in lexical semantic processing (Rogers et al., 2004; Patterson, Nestor, & Rogers, 2007; Bi et al., 2011). At the same time, PTL activity in fMRI is modulated by factors related to verb argument structure (Shetreet, Palti, Friedmann, & Hadar, 2007; Thompson et al., 2007) and by the size of syntactic constituents (Pallier, Devauchelle, &

Dehaene, 2011). In light of this uncertainty as to the role of each region, we report results from all three regions, while focusing on the ATL as our primary region of interest.

Previous MEG studies indicate that the time window most likely to show combinatory effects in the ATL is 200–300 ms after the presentation of the verb (Bemis & Pylkkänen, 2011, 2013; Brennan & Pylkkänen, 2012). While some studies have demonstrated somewhat earlier sensitivity to syntactic properties of the stimulus (Hahne & Friederici, 1999; Pulvermüller, Shtyrov, Hasting, & Carlyon, 2008), these effects are likely to be related to form prediction (Dikker, Rabagliati, & Pylkkänen, 2009); there is no reason to assume that a verb's subcategorization distribution is reflected in its orthographic form. Both ATL and PTL show differential activity in sentences compared to word lists, both in the 200–300 ms and in the 300–400 ms time windows (Brennan & Pylkkänen, 2012). Lexical modulations are likely to be reflected in the M350/N400m component, which is evident in the 300–500 ms time region and shows sensitivity to frequency and lexical semantics (Halgren et al., 2002; Pylkkänen & Marantz, 2003; Pylkkänen, Llinás, & Murphy, 2006). This component encompasses most of the temporal lobe around 350 ms, and later spreads to the prefrontal cortex, including Broca's region (Halgren et al., 2002). Since there is some uncertainty as to the expected time window for each region, we analyze the activity in all three regions in three time windows: 200–300 ms, 300–400 ms and 400–500 ms.

4.2 Methods

4.2.1 Participants

18 participants (13 women) from New York City participated in the experiment. All subjects provided informed consent and were paid for their participation. Participants ranged in age from 20 to 44 (median 27). Two of the participants were excluded from analysis because of equipment failures (one man and one woman). All subjects were right-handed (assessed using the Edinburgh Handedness Inventory; Oldfield, 1971) and were native speakers of English with normal vision.

4.2.2 Stimuli

The verbs analyzed in this paper were part of a larger lexical decision study. All of the words presented in the experiment were monomorphemic and monosyllabic 4-letter English words. Many 4-letter words are ambiguous between verbs and nouns (e.g. *lock*). For the purposes of the present study, a verb is defined as a word that used as a verb at least twice as frequently as it is used as a noun, based on CELEX (Baayen & Piepenbrock, 1995). None of the words analyzed here can be used in any other part of speech (adjective, adverb, etc.). In total, there were 189 verbs, out of a total of 750 words presented in the experiment.

Nonwords were selected from the ARC nonword database (Rastle, Harrington, & Coltheart, 2002) such that there was no significant difference in mean bigram letter frequency between the word and nonword stimuli. Participants performed a short practice block, which consisted of 7 trials, during which they received feedback. The stimuli were subsequently presented in 15 blocks of 100 trials, in random order, such that in each sequence of ten trials, five were nonwords and five were words. Participants did not receive feedback on their answers.

4.2.3 Procedure

The experiment was conduced in the KIT/NYU facility at New York University. Prior to recording, the head shape of each participant was digitized to allow source localization and co-registration with structural MRIs (Fastscan; Polhemus, VT). We also digitized three fiducial points (the nasion and the left and right pre-auricular points) and the position of five coils, placed around the participants face. Once the participant was situated in the magnetically shielded room for the experiment, the position of these coils was localized with respect to the MEG sensors, allowing us to assess the position of the participant's head for source reconstruction. Data were recorded continuously with a 157-channel axial gradiometer (Kanazawa Institute of Technology, Kanazawa, Japan) in a dimly lit magnetically shielded room.

Stimuli were presented using PsychToolBox (Pelli, 1997; Brainard, 1997) and projected onto a screen approximately 50 cm away. The stimuli were presented in white 30-point Courier font,

on a gray background. Each trial began with a fixation point (+) that appeared on the screen for 300 ms, followed by a blank screen for 300 ms, after which the stimulus was presented for 300 ms. Subjects then responded to the stimulus by pressing one of two buttons with their left hand to indicate whether or not they recognized the stimulus as a word. If the subject did not respond within 2 seconds, the next word was presented (this only happened in 5 trials in one of the subjects). The inter-trial interval was randomly selected between 300 ms and 600 ms (in 50 ms increments).

4.2.4 Data processing

The preprocessing and analysis of the MEG data closely followed the procedures of Solomyak and Marantz (2009a, 2009b). Environmental noise was removed from the data by regressing signals recorded from three orthogonally oriented magnetometers approximately 20 cm away from the recording array against the recorded data using the continuously adjusted least squares method (CALM; Adachi, Shimogawara, Higuchi, Haruta, & Ochiai, 2001). The data were then low-pass filtered to 40 Hz, resampled to 250 Hz to facilitate analysis, and high-pass filtered at 0.1 Hz. MEG channels in which there was no signal or excessive amounts of noise were interpolated from neighboring channels or rejected (at most 3 per subject). Trials in which at least one channel showed a peak-to-peak amplitude exceeding 3000 fT were rejected (the number of rejected trials ranged from 74 to 497, mean 150, median 106).

4.2.5 Source space analysis

MNE software (Martinos center MGH, Boston) was used to estimate neuroelectric current strength based on the recorded magnetic field strengths using minimum l_2 norm estimation (Dale & Sereno, 1993; Hämäläinen et al., 1993). Current sources were modeled as three orthogonal dipoles spaced approximately 5 mm apart across the cortical surface (Dale et al., 2000), yielding approximately 2500 potential electrical sources per hemisphere. For nine of the 16 subjects, structural MRIs were available from previous experiments, and their cortical surface was reconstructed based on their structural MRI using Freesurfer (Martinos center). For the 7 remaining subjects, a cortical surface

based on an averaged brain provided by Freesurfer was used. The neuromagnetic data was co-registered with the structural MRI (9 subjects) or the averaged cortex (7 subjects) using MNE by first aligning the fiducial points, and then using an Iterative Closest Point algorithm to minimize the difference between the points defining the head shape of each participant, and the scalp.

The forward solution was calculated for each source using a single-layer boundary element model (BEM) based on the inner-skull boundary. The estimated activation was normalized by dividing the estimated activation by the predicted standard error of the estimate, yielding Dynamic Statistical Parametric Maps (Dale et al., 2000).

Regions of interest were defined anatomically, using on the cortical parcellation performed by FreeSurfer based on the Desikan-Killiany gyral atlas (Desikan et al., 2006). The middle and superior temporal gyri were manually divided into anterior and posterior portions, using the anterior edge of the transverse temporal gyrus as a dividing landmark (following Brennan & Pylkkänen, 2012). Signed activity was summed across each ROI.

4.3 Lexical variables

Subcategorization entropy and relative entropy were calculated based on the definitions listed above. Subcategorization frame frequencies were obtained from the automatically acquired subcategorization lexicon VALEX (Korhonen, Krymolowski, & Briscoe, 2006). We used their Lexicon 5, a filtered version of the lexicon, which only includes frames whose relative frequency is higher than a frame-specific frequency, or frames that are listed in the manually created ANLT (Boguraev & Briscoe, 1987) and COMLEX (Grishman, Macleod, & Meyers, 1994) dictionaries. Before the filtering step, the subcategorization frame distributions were smoothed using linear interpolation. We used the ANLT subcategorization frame typology, which distinguishes 28 different frames in total.

As a control variable, we also replicated the contextual distinctiveness (CD) variable propose by S. A. McDonald and Shillcock (2001). This variable measures the extent to which the collocational

context of a word diverges from the average collocational context in the language. Consider, for example, the words *customer* and *lane*, which have identical frequency, yet differ in their CD: *lane* has a CD of approximately 1 bit, whereas *customer* has CD of approximately 0.5 bit. This reflects the fact that *lane* occurs in several common collocations (*fast lane*, *bike lane*), and therefore diverges more than *customer* from the average collocational context (S. A. McDonald & Shillcock, 2001). Disfluencies such as *ah* and *erm* receive the lowest CD values (very close to 0 bits), because they can occur in any context. At the opposite end of the CD spectrum are words such as *amok*, which only occur in a specific collocation (*run amok* in this case).

Collocational context was calculated from the full British National Corpus. We removed punctuation, capitalization and sentence boundary information. The corpus was lemmatized using the WordNet lemmatizer included in the Python Natural Language Toolkit (NLTK; Bird, Klein, & Loper, 2009), taking into account the part-of-speech tagging provided with the corpus. Frequent function words, such as pronouns and common prepositions, were removed based on NLTK's "stop word" list.[2] CD has two free parameters: the number of context words (i.e., the size of the vector representation), and the size of the window around each target word. We selected the values reported by McDonald and Shillcock to be optimal: 500 content words (chosen to be the most frequent words in the corpus), and a window of five words on either side of the target word.

More formally, the prior distribution $P(c_i)$ is defined as the overall distribution of context words in the corpus, independent of the target word. The posterior distribution $P(c_i|w)$ is the distribution of context words around the target word w. The CD of a target word w is then defined as the relative entropy between the prior distribution $P(c_i)$ and the posterior distribution $P(c_i|w)$:

[2]In following McDonald and Shillcock's (2001a) methodology in computing Contextual Distinctiveness, we remove from consideration close connections between verbs and prepositions (e.g., *depend on*) and verbs and particles (e.g., *look up*). Since in these cases the verb predicts a particular preposition or particle in its immediate syntactic environment, these dependencies may fall under the same kind of knowledge for verbs as subcategorization frames. Alternatively, or they might pattern with knowledge of collocational context. The nature of these dependencies should be the topic of further research.

	Entropy	Relative entropy	Frequency	CD	Number of senses
Entropy		-0.12	0.24	-0.07	-0.08
Relative entropy			0.15	-0.14	-0.15
Frequency				0	0
CD					-0.18

Table 4.1: Pearson correlations between lexical variables (after regressing log frequency out of CD and number of senses)

$$\text{CD}(w) = D(P(c)|P(c|w)) = \sum_{i=1}^{n} P(c_i|w) \log_2 \frac{P(c_i|w)}{P(c_i)} \tag{4.3}$$

In addition to contextual distinctiveness, we controlled for the verbs' number of senses, as listed in WordNet (Miller, 1995). This was done to address the concern that a larger number of senses may be correlated with higher SCF entropy, if the different senses of the verb select different frames (Roland & Jurafsky, 2002; Hare, McRae, & Elman, 2003).

The final control variable was log-transformed frequency, as listed in the SUBTLEX database (Brysbaert & New, 2009). Log-transformed frequency was correlated with both number of senses ($r = 0.53$) and contextual distinctiveness ($r = -0.52$). To reduce collinearity, we regressed frequency out of both variables. The residualized variables were highly correlated with the original variables (number of senses: $r = 0.84$, contextual distinctiveness: $r = 0.85$), suggesting that they can be interpreted in the same way. Following this residualization step, correlations between variables were all mild ($r < 0.25$): see Table 4.1.

4.4 Results

4.4.1 Behavioral

The accuracy of the subjects' responses ranged from 83.6% to 98.4% (mean 92.8%, median 93.5%). Mean reaction times ranged from 498 ms to 984 ms (mean 671 ms, median 644 ms).

Reaction times were log transformed and submitted to a linear mixed effects model (Bates et al., 2014). We included a by-item intercept, a by-subject intercept, and a by-subject slope for the two subcategorization variables that were of main interest in this paper. We did not include by-subject slopes for the control variables, because models with more elaborate random effect structures often did not converge.

Table 4.2 shows the model fitted to the verb trials. The p-values for fixed effects here and in what follows are derived using model comparison: the full model is compared to a model with the same random effect structure but without the predictor for which the p-value is being calculated. The difference in log likelihood between the partial and full model is then evaluated using the χ^2 approximation: $-2LL \sim \chi^2(1)$.

Only frequency had a significant effect on reaction times in verb trials. The effect went in the expected direction: Frequent verbs were responded to faster. SCF entropy did not affect reaction times. SCF relative entropy and contextual distinctiveness both showed non-significant trends in the expected direction: Verbs with an unusual SCF distributions or unusual collocational contexts elicited somewhat longer reaction times ($p = 0.12$ and $p = 0.11$, respectively).

An additional model was fitted to the entire data set, including nouns (Table 4.3). The subcategorization variables were excluded from the analysis, because they were not applicable to nouns. By-subject random slopes for frequency and contextual distinctiveness were added to the model. The effect of frequency was again highly significant ($p < 0.001$). Contextual distinctiveness also reached significance in the larger data set ($p = 0.004$): Words with more unusual contexts were responded to more slowly, replicating S. A. McDonald and Shillcock (2001).

4.4.2 MEG results

We analyzed the total neural activity in each of the three left-hemisphere regions of interest—anterior temporal lobe (ATL), posterior temporal lobe (PTL) and Broca's area. Activity was averaged in three 100 ms time windows: 200–300 ms, 300–400 ms and 400–500 ms after stimulus onset, based on the timecourse of effects in previous MEG studies (Brennan & Pylkkänen, 2012;

Predictor	Estimate	Std. Error	t-value	p-value (χ^2)
SCF entropy	0.0005	0.018	0.03	0.92
SCF relative entropy	0.018	0.017	1.11	0.12
Frequency	-0.018	0.003	-5.6	< 0.001
Contextual distinctiveness	0.04	0.03	1.38	0.11
Number of senses	0.001	0.0008	1.25	0.18

Table 4.2: Linear mixed-effects model fit to reaction times (verbs only)

Predictor	Estimate	Std. Error	t-value	p-value (χ^2)
Frequency	-0.022	0.002	-12.4	< 0.001
Contextual distinctiveness	0.028	0.01	2.55	0.004
Number of senses	0.006	0.0006	0.89	0.247

Table 4.3: Linear mixed-effects model fit to reaction times (all words)

Bemis & Pylkkänen, 2011). A linear mixed-effects model was fitted to each time window in each ROI.

4.4.2.1 Verbs

Subcategorization entropy was negatively correlated with ATL activity between 200 ms and 300 ms ($\beta = -0.063, p_{\chi^2} = 0.009$): Higher entropy verbs elicited less ATL activity. This correlation was weaker and no longer significant between 300 ms and 400 ms ($\beta = -0.045, p_{\chi^2} = 0.11$). Also between 300 ms and 400 ms, there was a marginal positive correlation with subcategorization relative entropy, such that higher relative entropy resulted in increased activity ($\beta = 0.052, p_{\chi^2} = 0.06$). Since the two variables are slightly correlated ($r = -0.12$), p-values derived from model comparison are somewhat conservative, suggesting that the marginal effect may well be reliable. The two subdivisions of the ATL, the aMTG and the aSTG, did not differ in the qualitative pattern of results.

PTL activity showed no effect of SCF entropy, and a marginal effect of SCF relative entropy between 300 ms and 400 ms ($\beta = -0.035, p_{\chi^2} = 0.08$). This marginal effect went in the opposite direction than the ATL effect: higher relative entropy resulted in less PTL activity. There was additionally a marginal effect of frequency, such that higher frequency words elicited more PTL activity (300–400 ms: $\beta = 0.008, p_{\chi^2} = 0.06$, 400–500 ms: $\beta = 0.009, p_{\chi^2} = 0.08$). An inspection of the two subdivisions of the PTL showed that the relative entropy effect was primarily in the pMTG (300–400 ms: $\beta = -0.052, p_{\chi^2} = 0.029$), and the frequency effect was stronger in the pSTG (300–400 ms: $\beta = 0.01, p_{\chi^2} = 0.03$; 400–500 ms: $\beta = 0.012, p_{\chi^2} = 0.03$).

Neither of the SCF variables had a significant effect in Broca's area. Overall, the SCF entropy effect was specific to the ATL: None of the other regions showed a significant effect of this variable, and an ANOVA revealed a significant interaction between SCF entropy and region ($p_{\chi^2} = 0.012$).

4.4.2.2 All words

We also analyzed the entire set of words. Due to model convergence issues, we fitted separate models for each of the three relevant variables—frequency, contextual distinctiveness and number of senses—each with the relevant by-subject slope. Since the variables were decorrelated from each other, the results should be very similar to a model containing all of the variables.

There were no effects of contextual distinctiveness. Number of senses had a marginal negative effect in Broca's area between 300 ms and 400 ms ($\beta = 0.002, p_{\chi^2} = 0.05$). Frequency, on the other hand, had a significant effect between 300 ms and 400 ms after simulus presentation, both in the PTL ($\beta = 0.008, p_{\chi^2} = 0.0007$) and in Broca's area ($\beta = 0.009, p_{\chi^2} = 0.0005$). In both areas, more frequent words led to increased activity. The effect of frequency between 300 ms and 400 ms in the ATL did not reach significance. However, it trended in the same direction as the frequency effect in other regions ($\beta = 0.005, p_{\chi^2} = 0.08$), and a direct comparison between the effects of frequency in the ATL and the PTL did not reveal a significant frequency by region interaction ($p_{\chi^2} = 0.4$).

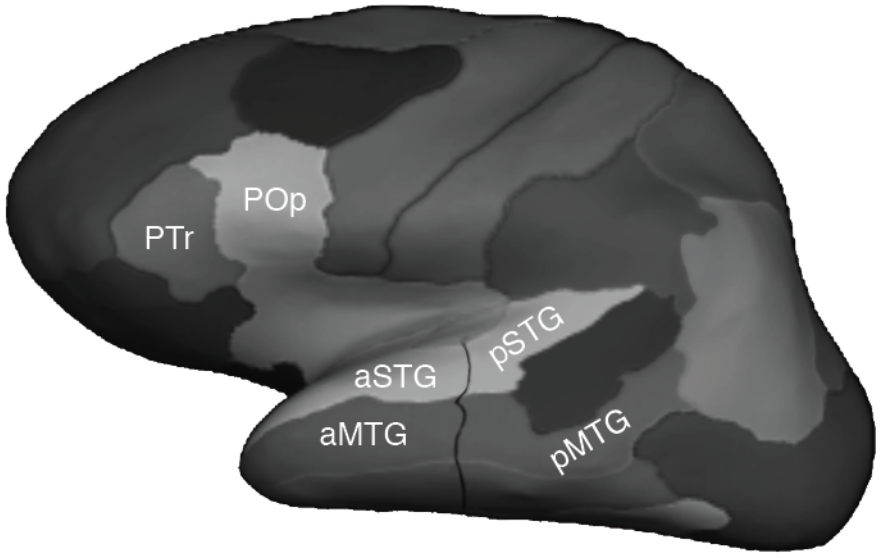


Figure 4.1: Language-related regions of interest in the left hemisphere. 1. Broca's area: PTr – inferior frontal gyrus, pars triangularis (Brodmann area 45); POp – inferior frontal gyrus, pars opercularis (Brodmann area 44). 2. Anterior temporal lobe (ATL): aMTG – anterior middle temporal gyrus; aSTG – anterior superior temporal gyrus. 3. Posterior temporal lobe (PTL): pSTG – posterior superior temporal gyrus; pMTG – posterior middle temporal gyrus.

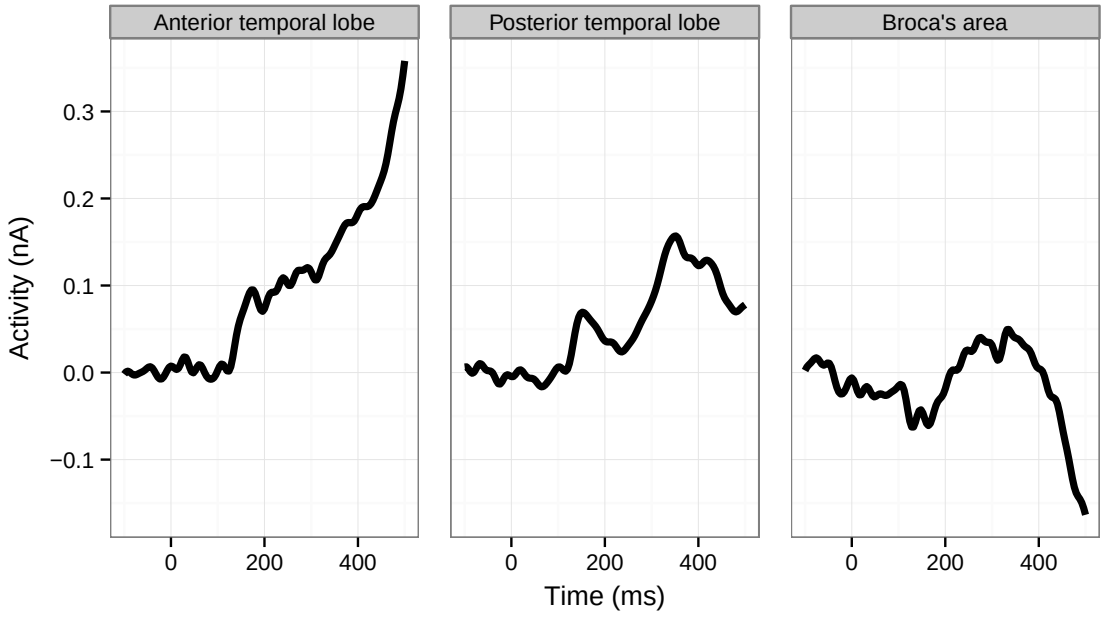


Figure 4.2: Verb trials: grand mean of neural activity, across subjects

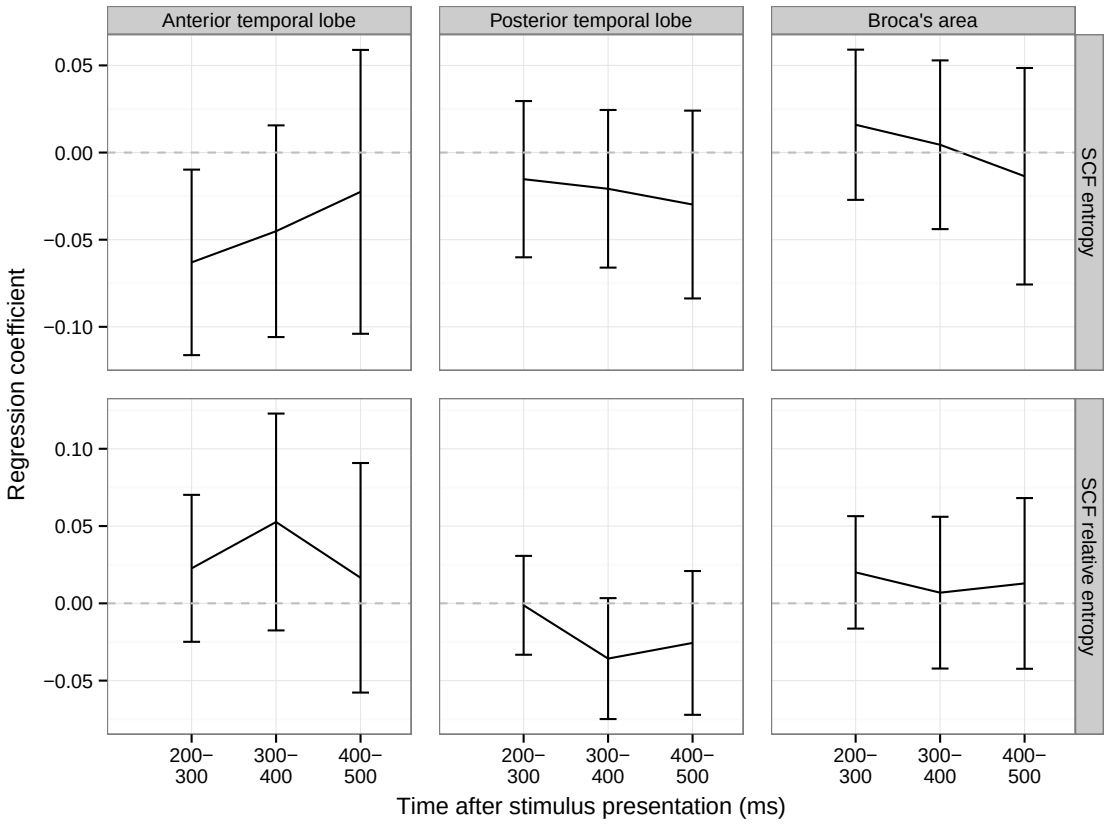


Figure 4.3: Verb trials: Regression coefficients for subcategorization variables. Approximate confidence intervals are calculated as twice the standard error of the regression coefficients in each direction.

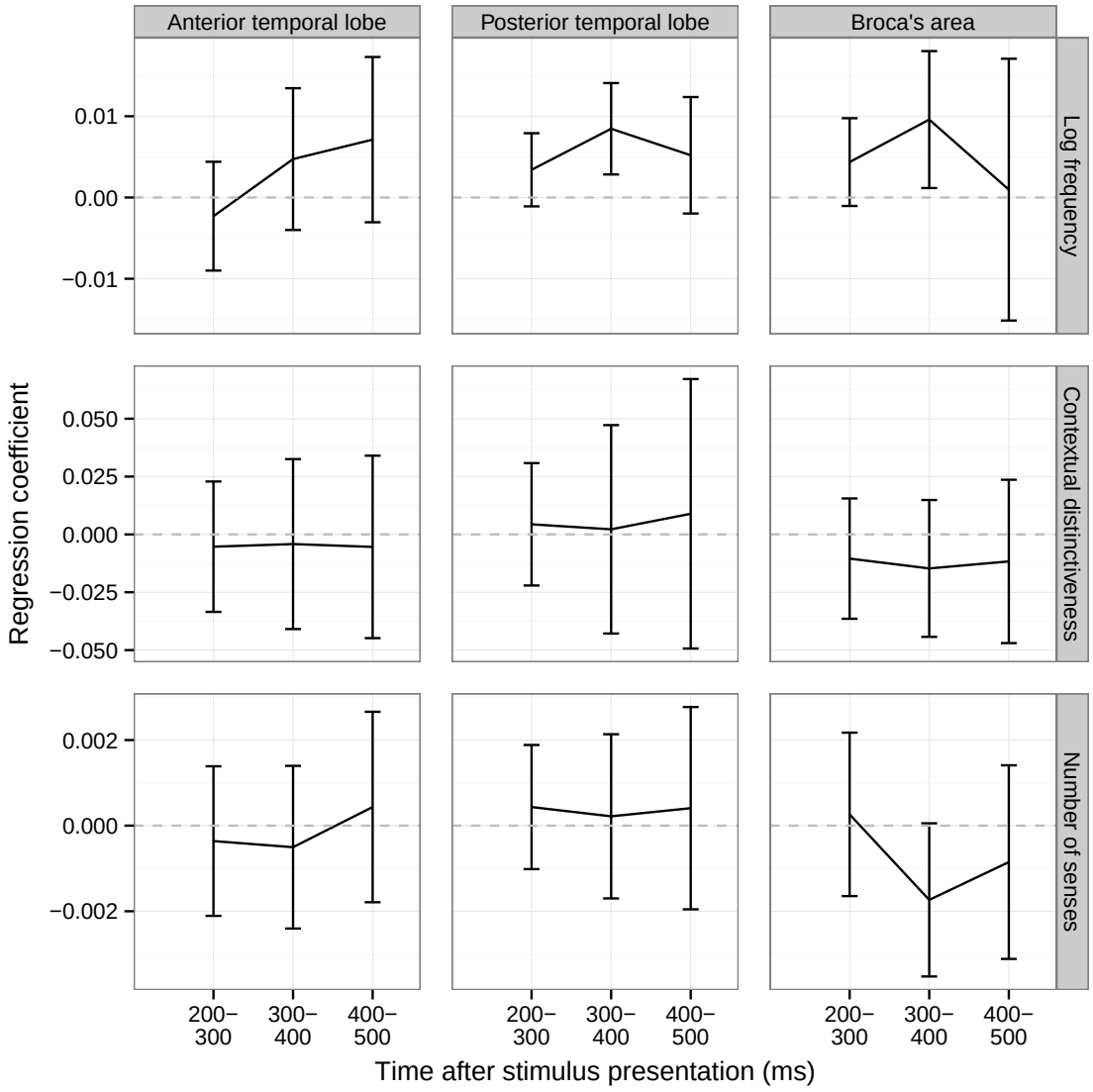


Figure 4.4: All trials: Regression coefficients for control variables. Approximate confidence intervals are calculated as twice the standard error of the regression coefficients in each direction.

4.5 Discussion

The present study showed that the typical syntactic context of a word—specifically, the distribution of a verb's subcategorization frames—affects activity in the left anterior temporal lobe (ATL) during the recognition of the word. Verbs with high subcategorization frame entropy elicited less ATL activity than low entropy verbs between 200 ms and 300 ms, the time window associated with composition of phrases from individual words in the same region (Bemis & Pylkkänen, 2011). This effect could not be attributed to any of the control variables we investigated: word frequency, collocational context or number of senses.

The reduction in neural activity in response to increased entropy over continuations is in line with previous behavioral findings in the morphological domain. Morphological family entropy has a facilitatory effect in visual lexical decision (Baayen et al., 2006). Likewise, high morphological continuation entropy leads to shorter reaction times in auditory lexical decision (Baayen et al., 2007; Wurm et al., 2006). These parallels between word-internal morphological continuations and word-external syntactic continuations are expected in light of the similarities between word-internal and word-external context, as suggested by (Halle & Marantz, 1993; Baayen et al., 2011) (for a different view, see Cappelle, Shtyrov, & Pulvermüller, 2010; Pulvermüller, Cappelle, & Shtyrov, 2013). If competition entails more activity, as in the case of competition between the two meanings of an ambiguous word (Simon et al., 2012), the present pattern of results is at odds with an account whereby all possible continuations are activated and compete for selection: Such an account would predict higher activity for a larger number of continuations.

In addition to the entropy effect, there was a marginal effect of relative entropy in the ATL between 300 ms and 400 ms, such that higher relative SCF entropy caused an increase in activity. SCF relative entropy measures the divergence between the SCF distribution of the recognized verb and the average SCF distribution in the language. This result mirrors the effect of relative inflectional entropy: Serbian masculine words whose distribution over inflected forms diverges from that of the average masculine word take longer to recognize (Milin, Filipović Djurdjević, & Moscoso del Prado Martín, 2009; Baayen et al., 2011). The average SCF distribution is likely to

serve as the reader's prior distribution, in advance of seeing the specfic verb; the verb's specific distribution is more surprising the larger the divergence between this specfic distribution and the prior distribution. Increased neural activity in response to stimuli with unexpected properties ties in with theories of predictive coding (Rao, Ballard, et al., 1999; Friston, 2005), according to which neural responses reflect the degree to which the incoming stimulus forces an update in the reader's probabilistic expectations. Specifically in language comprehension, numerous studies have shown that words with unexpected properties elicit a stronger N400 component than expected words (see Kutas et al., 2011 for a recent review). A potential interpretation of the later latency of the marginal relative SCF entropy effect (300–400 ms) compared to the SCF entropy effect (200–300 ms) may be that the relative entropy effect is part of the N400 surprise response, whereas the entropy effect reflects an earlier combinatorial component (Bemis & Pylkkänen, 2011).

Although the ATL was our main region of interest, we tested two additional left-hemisphere regions as controls: the posterior temporal lobe (PTL) and Broca's area. The two control regions did not show an SCF entropy effect. Conversely, the two regions showed significant frequency effects between 300 ms and 400 ms, in contrast with the ATL. Increased activity in response to frequent words has been observed in previous studies (Yarkoni, Speer, Balota, McAvoy, & Zacks, 2008; Solomyak & Marantz, 2009a; Brennan et al., 2012). These results are in line with models that ascribe a compositional role to the ATL and a lexical role to the PTL (Hickok & Poeppel, 2007). This potential dissociation between the anterior and posterior parts of the temporal lobe should be interpreted with caution, however. While an interaction test showed that the two control regions differed significantly from the ATL in the magnitude of the subcategorization entropy effect, a similar test did not show a significant difference in the effect of frequency. Furthermore, there was a hint of an effect of subcategorization relative entropy in the PTL between 300 ms and 400 ms, and fMRI studies have found effects of verb argument structure in posterior temporal regions (Shetreet et al., 2007; Shetreet, Friedmann, & Hadar, 2010).

The facilitatory effect of continuation entropy may reflect a conservative prediction strategy, whereby a potential continuation is only preactivated if there is low uncertainty as to the identity

of that continuation. Dikker and Pylkkänen (2013), for example, observe increased neural activity in constraining contexts relative to non-constraining contexts, in advance of the presentation of the predictable word. This interpretation of the facilitatory effect of uncertainty over continuations predicts that the effect is not specific to individual word recognition tasks, and should also show up when the verb is embedded in a sentence. In a sentential context, if a verb licenses a specific syntactic prediction, that syntactic category is immediately activated. Otherwise, the parser waits for the complement to establish the syntactic structure of the verb phrase.

An alternative interpretation of the results would be that words that are encountered in multiple different contexts are more robustly represented in the brain than words that always occur in one specific context (Adelman, Brown, & Quesada, 2006). This interpretation would not be compatible with a strong functional interpretation of the anatomical pattern of results: subcategorization entropy only modulated activity in the anterior part of the temporal lobe, whereas it is the posterior part which is most strongly associated with the retrieval of lexical representations. The two interpretations sharply diverge in their prediction for head-final languages, in which verbs do not typically predict the syntactic structure of their arguments: the conservative prediction hypothesis predicts that the subcategorization entropy effect will be weaker or non-existent in head-final languages, whereas the robustness of representation hypothesis does not predict any difference between head-final and head-initial languages.

The results of this study suggest that syntactic information associated with a word is accessed even when structure building is not required by the experimental task. Since words are almost always encountered in context, it is not surprising that the activation of syntactic information is automatic. The subcategorization biases of a single verb prime have been shown to affect subjects' productions in a syntactic priming paradigm (Melinger & Dobel, 2005), and argument-structure based distinctions modulate brain activity in single-word lexical decision (Thompson et al., 2007).

Neither of our subcategorization frame variables affected lexical decision reaction times. This may be due to insufficient power: while there was a robust effect of contextual distinctiveness on the full set of words (including nouns), it did not reach significance on verbs. This indicates that

our set of verbs may have been too small to detect behavioral effects of contextual variables. In addition, it is possible that syntactic information, while automatically activated upon reading a word, is not recruited to make lexical decisions. Previous studies have also failed to detect effects of the number of subcategorization frames on reaction times in lexical decision (Schmauder, 1991; Shapiro, Zurif, & Grimshaw, 1987). It is not clear why that should be the case, and further research would be necessary to address this point.

Subcategorization frame entropy is not a perfect measure of a verb's typical syntactic context. Verbs also differ in their likelihood of occurring with different classes of adjuncts (*yesterday*, *with a knife*). Furthermore, verbs, and words in general, may vary in the larger syntactic structures in which they tend to appear – e.g., questions, ellipsis constructions and so on. It is possible that subcategorization frames are more tightly connected to the lexical representation of a verb than other types of syntactic context (Boland, 2005). It thus remains an open question whether syntactic context will always have the same effect as subcategorization frames in word recognition.

In conclusion, this study reinforces the role of a word's typical contexts on its recognition in isolation, and in particular the role of the abstract syntactic context, over and above the specific lexical items that tend to cooccur with the word (collocational context). We found that verbs that tend to appear with a larger variety of syntactic arguments, as measured by subcategorization frame entropy, elicited less neural activity in the left anterior temporal lobe. This is the opposite of what would be predicted by a competition account, under which the activation of multiple possible frames should lead to larger overall activity. We hypothesized that this effect reflects a conservative prediction strategy: a syntactic frame is only preactivated when the verb licenses a specific prediction.

5 Competition and prediction in complex spoken words

5.1 Introduction[1]

Language comprehension is predictive: as soon as language users any kind of information about the utterance they are processing, they use it to make predictions about the way the rest of the utterance will unfold (Altmann & Mirković, 2009). Words that are more predictable are read faster and elicit weaker neural responses (Ehrlich & Rayner, 1981; Kutas & Hillyard, 1984), likely because words are predicted in advance of being encountered (DeLong et al., 2005; Fruchter et al., 2015). Word predictability effects have also been reported in spoken word processing (Van Petten, Coulson, Rubin, Plante, & Parks, 1999).

Word predictability effects fit into the broader neuroscientific framework of predictive coding (Rao et al., 1999; Friston, 2005). In this framework, higher level brain regions generate predictions and propagate them to lower level sensory cortex; neural activity in sensory regions reflects prediction error. Evidence for this framework has been reported in the visual modality (Summerfield & Egner, 2009; Dikker et al., 2009) as well as in auditory percetion (Sohoglu, Peelle, Carlyon, & Davis, 2012; Todorovic & de Lange, 2012; Wacongne, Changeux, & Dehaene, 2012). In partic-

[1]This work was done in collaboration with Laura Gwilliams, Phoebe Gaston and Alec Marantz. An early version was presented as a poster at the Society for the Neurobiology of Language, Amsterdam, September 2014.

ular, there are recent reports that less expected speech segments in newly learned nonwords elicit increased neural activity in temporal regions associated with speech processing (Gagnepain et al., 2012), and that this effect extends to real words (Ettinger et al., 2014).

5.1.1 Morphological prediction and competition

The present study highlights an intermediate level of analysis between the segment level and the word level. Many words have internal structure; the English word *builder*, for instance, can be viewed as composed of two units, called *morphemes*: the stem *build* and the suffix *-er*. In languages such as Turkish virtually all words are composed of multiple morphemes. Much experimental work has shown that morphological structure plays an important role in visual and auditory word recognition (Marslen-Wilson, Tyler, Waksler, & Older, 1994; Solomyak & Marantz, 2009a; Taft & Forster, 1975). As linguists have observed, there is often little principled reason to distinguish complex words like *builder* from multiword sequences like *cable guy* (Di Sciullo & Williams, 1987; Halle & Marantz, 1993); this is in particular the case in auditory word recognition, which of course does not include spaces between words.

The lack of principled distinction between word-internal and word-external regularities raises the possibility that listeners predict not only at the level of individual segments and at the level of whole words, but also at the intermediate morphological level: the next morpheme of the word may be predicted from the morphemes heard so far. Supporting this hypothesis is a recent study that investigated the neural activity in auditory cortex corresponding to the second syllable of in bimorphemic words (*builder*) compared to bisyllabic monomorphemic words (*bourbon*) (Ettinger et al., 2014). Both classes contained words with a predictable second syllable as well as words with an unpredictable second syllable. Ettinger et al. (2014) found that predictability effects on the second syllable were stronger in bimorphemic than in monomorphemic words. This suggests that listeners were using the first morpheme (*build*) to predict the second one (*-er*).

A separate line of work, mostly in behavioral auditory lexical decision experiments, points to the existence of morphological competition effects: Words whose first morpheme is shared with

many other words are typically recognized more slowly (Balling & Baayen, 2012; Kemps et al., 2005; Meunier & Segui, 1999; Wurm et al., 2006). These findings are taken to suggest that all suffixed words that begin with a given stem are activated when the stem is perceived.

5.1.2 The current study

The present study has two goals: first, to replicate the recently reported effects of speech segment and morpheme prediction error; and second, to combine the two strands of research on morphological processing in spoken word recognition by exploring how competition among potential upcoming morphemes affects morphological prediction error. Participants listened to isolated bimorphemic words while their brain activity was recorded using MEG; some of the words were followed by a comprehension question. The words varied both in the predictability of the second morpheme from the first one and in the amount of morphological competition at the morpheme boundary.

A straightforward effect of morphological prediction error will manifest in increased neural activity in response to less expected suffixes. If neural activity is affected by the range of suffixes that can be predicted at the stem, we expect higher suffix competition to result in increased neural activity during the processing of the first morpheme. Finally, competition may modulate the effect of predictability at the suffix, by making prediction more conservative when competition is high. In other words, listeners may be more commited to their predictions in low competition scenarios, in which case an unexpected suffix may be more surprising in low competition than in high competition cases (Ettinger et al., 2014; Linzen & Jaeger, 2014).

5.2 Method

5.2.1 Calculation of predictors

5.2.1.1 Phonological surprisal

In calculating segment surprisal, we follow the standard assumption of the cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), according to which speech segments are perceived and processed veridically and sequentially. The set of words that are compatible with the segments perceived so far is referred to as the cohort. As soon as a segment is perceived that is incompatible with one of the words in the cohort, that word drops out of the cohort. Not all members of the cohort are equally likely: In the absence of context, a reasonable listener will assume that frequent words are more likely than infrequent ones (Norris, 2006). Word frequency in the language therefore defines a probability distribution over the cohort, which in turn induces a probability distribution over the next segment: The probability that a given segment will be the next segment is the sum of the probabilities of all of the words in the cohort that have this segment as their next segment.

Formally, let C be a function that maps a sequence of segments to the set of all words that start with that sequence; e.g.,

- $C(\langle \text{k}, \text{ə} \rangle) = \{$*corrupt, communicate, computer, complain, collision,* $\ldots\}$
- $C(\langle \text{k}, \text{ə}, \text{m}, \text{p} \rangle) = \{$*complain, computer,* $\ldots\}$

Let $f(w)$ be the frequency of the word w and $F(S)$ the summed frequency of all words in a set S, i.e.

$$F(S) = \sum_{w \in S} f(w) \tag{5.1}$$

The probability distribution of the i-th segment conditioned on the first $i-1$ segments is then given by

$$P(s_i|s_1,\ldots,s_{i-1}) = \frac{F(C(\langle s_1,\ldots,s_i\rangle))}{F(C(\langle s_1,\ldots,s_{i-1}\rangle))} \tag{5.2}$$

Building on evidence that the effect of predictability on reading times is logarithmic (Smith & Levy, 2013), we hypothesize that prediction error E at a segment will be proportional to its surprisal, or the inverse of the logarithm of its conditional probability given the segments received so far (Ettinger et al., 2014):

$$I(s_i|s_1,\ldots,s_{i-1}) = -\log_2 P(s_i|s_1,\ldots,s_{i-1}) \tag{5.3}$$

5.2.1.2 Morphological predictors

We operationalize the competition after the first morpheme as the entropy of the word's morphological cohort after that morpheme. The morphological cohort is defined by analogy to the phonological one. After the first morpheme (e.g., *build-*) has been processed, the morphological cohort includes all of the words that have this morpheme as their first morpheme (e.g., *builders*, *building*). We define the function $C_M(\langle m_1,\ldots,m_k\rangle)$, which maps a sequence of morphemes to the corresponding morphological cohort:[2]

- $C_M(\langle$*build*$\rangle) = \{$*build, builder, builders, building,* $\ldots\}$
- $C_M(\langle$*build, -er*$\rangle) = \{$*builder, builders*$\}$

Let $H(S)$ be the entropy of the probability distribution defined by the frequencies of the members of a set S, i.e.

$$H(S) = -\sum_{w\in S} \frac{f(w)}{F(S)} \log_2 \frac{f(w)}{F(S)} \tag{5.4}$$

[2]See next section for a discussion of whether or not the bare stem *build* should be in the cohort.

Our measure of competition at the morpheme boundary is the morphological cohort entropy of a stem m_1, i.e. $H(C_M(\langle m_1 \rangle))$. The definition of the surprisal of the second morpheme given the stem is analogous to the definition of segment surprisal:

$$I(m_2|m_1) = -\log_2 \frac{F(C_M(\langle m_1, m_2 \rangle))}{F(C_M(\langle m_1 \rangle))} \tag{5.5}$$

5.2.1.3 Inclusive and exclusive morphological predictors

As mentioned in the previous section, after the first morpheme (e.g., *build-*) has been processed the morphological cohort includes all of the words that have this morpheme as their first morpheme (e.g., *builders*, *building*). This definition is ambiguous as to whether or not the bare stem (*build*) is included in the cohort. There is evidence that listeners can distinguish monosyllabic words (*cap*) from longer words that start with the same syllable (*captain*), even before the end of the first syllable; these phonetic cues may be strong enough to stop listeners from activating *captain* when they hear *cap* (Davis, Marslen-Wilson, & Gaskell, 2002). Likewise, the subphonemic detail in *build-* as a first morpheme of a multimorphemic word may be sufficient for listeners to rule out the word *build* (Blazej & Cohen-Goldberg, 2014). It may therefore be useful to define two separate types of morphological cohorts: an inclusive cohort that includes the bare stem and an exclusive cohort that doesn't. Correspondingly, we can define *two* functions, $C_M^{inc}(\langle m_1, \ldots, m_k \rangle)$ and $C_M^{exc}(\langle m_1, \ldots, m_k \rangle)$, which map sequences of morphemes to the sets of words that include them:

$$\begin{aligned}
C_M^{inc}(\langle build \rangle) &= \{build, builder, builders, building, \ldots\} \\
C_M^{exc}(\langle build \rangle) &= \{builder, builders, building, \ldots\} \\
C_M^{inc}(\langle build, -er \rangle) &= \{builder, builders\} \\
C_M^{exc}(\langle build, -er \rangle) &= \{builders\}
\end{aligned} \tag{5.6}$$

We can now define the inclusive entropy of a stem m as $H(C_M^{inc}(\langle m \rangle))$, and its exclusive entropy as $H(C_M^{exc}(\langle m \rangle))$. Similarly, there are two ways to define surprisal, depending on whether or not the bare stem remains in the cohort. The inclusive surprisal of an affix m_2 given the stem m_1 is defined as above:

$$I^{inc}(m_2|m_1) = -\log_2 \frac{F(C_M^{inc}(\langle m_1, m_2 \rangle))}{F(C_M^{inc}(\langle m_1 \rangle))} \tag{5.7}$$

Whereas the exclusive surprisal is defined as

$$I^{exc}(m_2|m_1) = -\log_2 \frac{F(C_M^{inc}(\langle m_1, m_2 \rangle))}{F(C_M^{exc}(m_1))} \tag{5.8}$$

Note that the numerator makes reference to the *inclusive* morphological cohort starting with m_1 and m_2, because we assume that the point where the surprisal of m_2 has its effect would be closer to the beginning of this morpheme, where subphonemic detail distinguishing longer from shorter words may play a weaker role; however, this is another degree of freedom that should be systematically explored.

5.2.2 Materials

All of the target items were bisyllabic bimorphemic words extracted from the English Lexicon Project (Balota et al., 2007). All words had a stem + suffix structure, where the stem and the suffix each contained a single vowel and the suffix was derivational (as opposed to inflectional). None of the stems were bound roots, i.e., all of the stems could also serve as independent words (words such as *possible* or *tolerate* were excluded). None of the suffixes conditioned a sound change in the root; that is, items such as *facial* [feɪʃl] (from *face* [feɪs]) were excluded. We further excluded proper nouns and words that had an accuracy of 75% or lower in the English Lexicon Project lexical decision reaction time database. This was done to ensure that participants were familiar with all of the items.

Morphological segmentation was obtained from the English Lexicon Project. Frequency norms were obtained from the SUBTLEX-US corpus (Brysbaert & New, 2009). Surprisal and entropy are naturally correlated. Although our regression design allows us to incorporate correlated predictors, we selected the target words such that the correlations among log-transformed word frequency, surprisal and entropy were low, to maximize statistical power (all $|r| \leq 0.2$ between variables that were used in the same regression model). Exclusive and inclusive affix surprisal were naturally highly correlated ($r = 0.86$); interestingly, the correlation between inclusive and exclusive post-stem entropy was not as high ($r = 0.38$), suggesting that the relative frequency of the bare stem in the morphological family varies widely from stem to stem and is not a good predictor of the shape of the distribution of suffixed forms. Table 5.1 shows the full matrix of correlation coefficients between the variables; Figure 5.1 shows the detailed correlation between the various measures of cohort entropy and affix surprisal.

	Log-frequency	Inc. surp.	Exc. surp.	Inc. ent.	Exc. ent.
Log-frequency	1	-0.16	-0.11	0.12	0.12
Inclusive surprisal		1	(0.86)	-0.15	(0.2)
Exclusive surprisal			1	(0.25)	0.2
Inclusive entropy				1	(0.38)
Exclusive entropy					1

Table 5.1: Correlations between predictors. Values in parentheses indicate pairs of predictors that were not used in the same analysis.

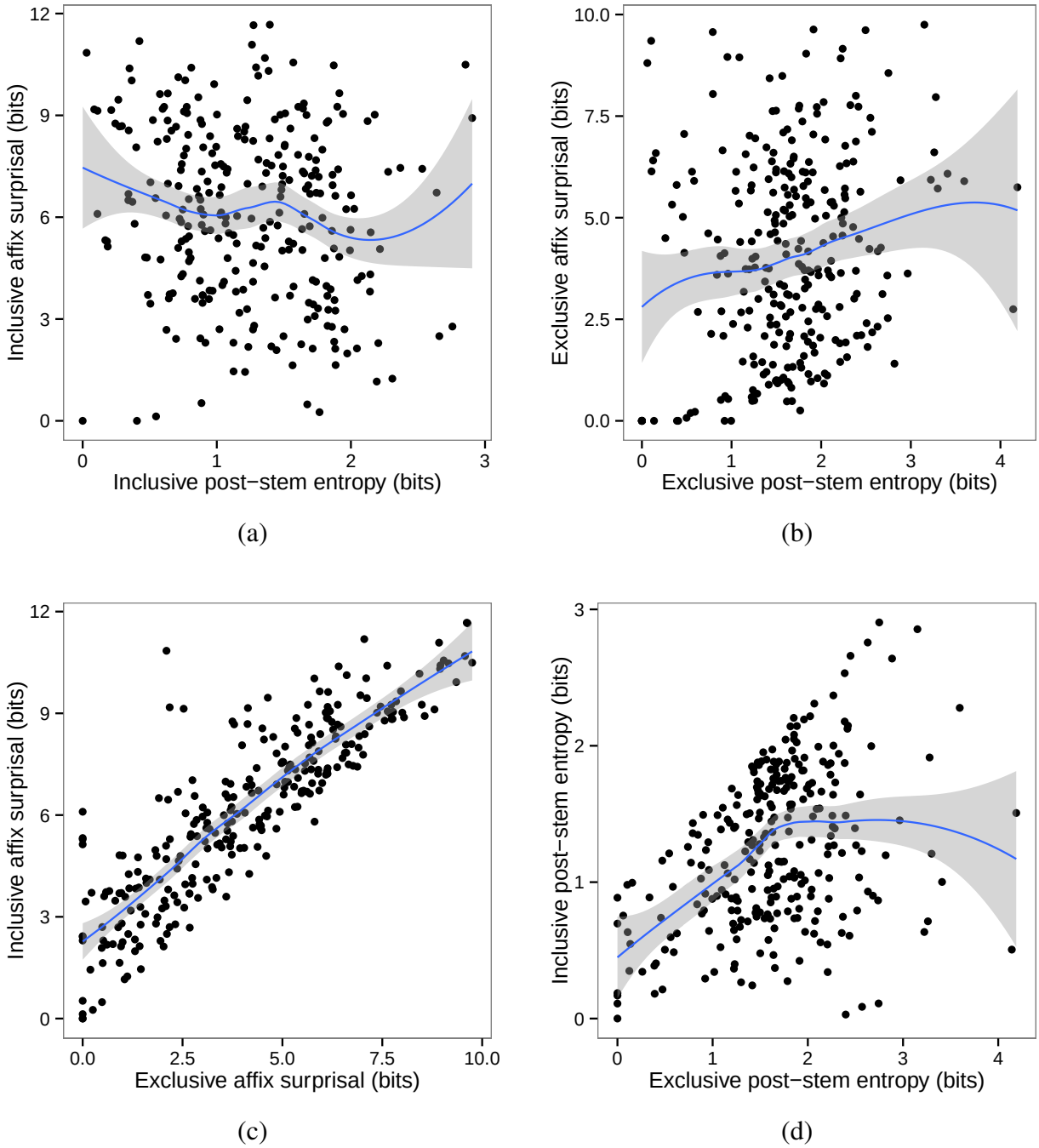


Figure 5.1: Detailed correlations between different measures of morphological entropy and surprisal. (a) and (b) show the correlations between variables that were used in the same regression model (either both exclusive or both inclusive variables). (c) and (d) show the correlation between the exclusive and inclusive versions of the same variable.

The experiment additionally included 100 filler items, designed to increase the range of morphological structures in the materials: 40 monosyllabic monomorphemic words (*fume*, *hoard*); 40 bisyllabic monomorphemic words (*tundra*, *varnish*); and 20 bisyllabic bimorphemic words with inflectional suffixes (*roughest*) or derivational suffixes attached to bound roots (*juncture*). All filler items had an accuracy of 75% or higher in the the English Lexicon Project lexical decision data. The mean frequency of the filler items was matched to that of the target items. None of the stems of the morphologically complex filler items appeared in any of the target items.

Paraphrases were created for 54 of the items (one in seven items): 40 for target items and 14 for filler items. Half of the paraphrases matched the item (*fizzy*: *full of bubbles*), and half did not (*hiker*: *a cyclist*). The paraphrases were designed to ensure that the participants were paying attention both to the stem (*smoky*: *full of water*) and to the suffix (*playful*: *a contestant*). None of the paraphrases contained words with the same stem as a target item.

Four experimental lists were randomly constructed such that trials with paraphrases were separated by at most 15 trials without paraphrases. To minimize potential order effects, each of these four lists was then reversed, for a total of eight experimental lists.

The materials were recorded by a male native English speaker. The mean duration of the sound files was 686 ms (standard deviation: 98 ms; median: 681 ms; range: 461–930 ms). Initial alignments between the transcriptions of the words in the CMU Pronouncing Dictionary and the acoustic stimuli were obtained from the Penn Forced Aligner (Yuan & Liberman, 2008) and were then manually corrected in Praat. A few discrepancies between the CMU Pronouncing Dictionary and the Unisyn transcriptions supplied with the English Lexicon Project were manually corrected. The location of the morpheme boundary was identified manually. The morpheme boundary was on average 327 ms after the beginning of the sound file (standard deviation: 72 ms; median: 320 ms; range: 182–541 ms). Figure 5.2 shows an example of an annotated waveform.

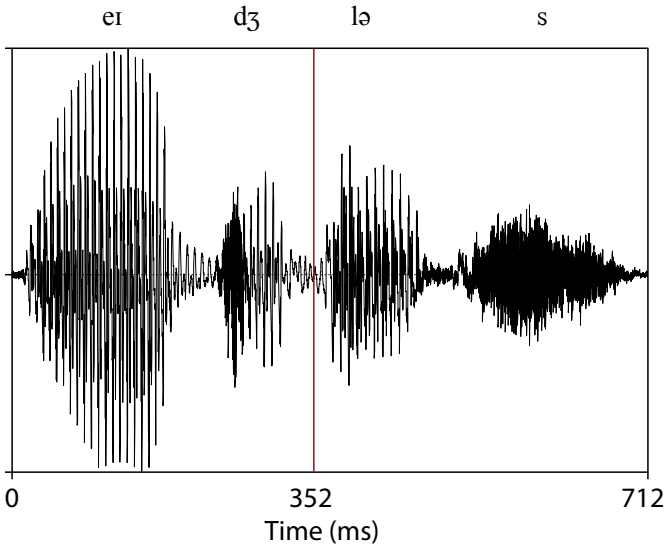


Figure 5.2: Annotated waveform for *ageless*. The vertical red line indicates the morpheme boundary.

5.2.3 Participants

Twenty-four native English speakers (11 female, ages 20-52, median age 22) were recruited from the NYU Abu Dhabi community. Handedness was assessed using the Edinburgh Handedness Inventory (Oldfield, 1971). All subjects provided written informed consent before participation. Three subjects were assigned to each of the eight lists.

5.2.4 Procedure

Stimuli were presented using Matlab PsychToolbox (Brainard, 1997; Pelli, 1997). The instructions were: "You will hear single words. After some of the words, you will hear a phrase. Press a button to indicate whether the phrase could replace the word in a sentence. When an asterisk shows up, press either of the keys when you are ready to move on to the next word."

The structure of each trial was as follows. A fixation cross was presented for 300 ms, followed by a 300 ms blank. The word was then played in the participants' headphones, followed by a 700 ms blank. In the trials that did not have a paraphrase, an asterisk was displayed until the participant pressed a button to move to the next trial. In the trials that had a paraphrase, a question mark was

displayed for 700 ms, followed by a 700 ms blank, after which the paraphrase was played. When the paraphrase finished playing, the probe "Accurate?" appeared on the screen and stayed on until the participant pressed a button.

Participants performed 10 practice trials outside the MEG machine, and 10 additional practice trials inside the machine. Feedback on the accuracy of the responses was provided after each trial. The trials were presented in four blocks of 95 trials each. At the end of each block, participants were instructed to press a key when they were ready to move to the next block.

5.2.5 Data acquisition and preprocessing

Prior to recording, the head shape of each participant was digitized using a dual source handheld FastSCAN laser scanner (Polhemus, VT, USA). We also digitized three fiducial points (the nasion and the left and right pre-auricular points) and the position of five coils, placed around the participant's face. The position of these coils was localized with respect to the MEG sensors, allowing us to assess the position of the participant's head for source reconstruction. Data were recorded continuously at the Neuroscience of Language Lab in NYU Abu Dhabi using a whole-head 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan) as subjects lay in a dimly lit, magnetically shielded room. Data were recorded at 1000 Hz and low-pass filtered at 200 Hz.

An off-line low-pass filter of 40 Hz was applied to data; the data was then downsampled at 100 Hz and split into 1200 ms epochs starting at the onset of each target word. Average baseline activity in the 100 ms preceding the onset of the epoch was subtracted from the epoch. Epochs with visible blink patterns were manually rejected (median proportion of rejected epochs: 20.5%).

The MNE-Python software package (Gramfort et al., 2014) was used to obtain distributed minimum norm source activity estimates. Since individual structural MRIs were not available, the forward solutions were based on the average cortex included in the FreeSurfer distribution, scaled to fit each individual participant's head shape to the extent possible. The cortical reconstruction was decimated into 10242 vertices in each hemisphere. Noise covariance matrices for each subject

were estimated based on the 100 ms pre-stimulus baseline period. The amplitudes of the dipoles at each source were normalized by the noise to form dynamic statistical parametric maps (dSPMs) (Dale et al., 2000). A separate source activity estimate was obtained for each trial. Given the absence of structural MRIs for our participants, the orientation of the dipoles was not constrained; this resulted in a three-dimensional vector at each source. The norm of this vector was taken as an estimate of the activity in that source.

Three epochs were defined for each trial, timelocked to three different anchors: (1) the onset of the stimulus (starting at the onset and ending 800 ms after it); (2) the morpheme boundary (200 ms before the boundary to 400 ms after it); and (3) the offset of the stimulus (200 ms before the offset to 300 ms after it).

5.2.6 Statistical analysis

5.2.6.1 Morphological prediction

For each subject, we performed a multiple linear regression analysis predicting single-trial dSPM values at each source and at each timepoint from the following variables: affix surprisal, morphological cohort entropy following the stem, log-transformed full word frequency, and the interaction between surprisal and entropy. All variables were centered before being entered into the regression. This resulted in four statistical parametric maps for each subject, one for each predictor. We then performed a spatiotemporal cluster permutation test for each predictor with a one-sample t-test, to detect contiguous regions in space and time in which regression coefficients were significantly different from zero at the group level (Maris & Oostenveld, 2007; Gramfort et al., 2014). We performed 10000 permutations per predictor; the criterion for inclusion in a cluster was an uncorrected t-statistic of 1.67.

5.2.6.2 Phonological prediction

When does the predictability of a segment start affecting auditory cortex neural activity? A natural option would be 100 ms after the onset of the segment. This lag reflects the latency of the M100 component and the analogous predictability-sensitive N1 EEG component (Astheimer & Sanders, 2011). Likewise, expectation effects in pure tone perception become evident around 100 ms after the tone (Todorovic & de Lange, 2012).

Assuming a fixed 100 ms lag between segment onset and the beginning of its neural correlates neglects the fact that the speech signal at any given moment is affected not only by the current segment but also by the segments surrounding it, due to the difficulty in coordinating different articulators in the vocal tract (coarticulation; see Farnetani & Recasens, 2013 for a review). In particular, the identity of an upcoming segment is often evident from its effect on the current segment (see Salverda, Kleinschmidt, & Tanenhaus, 2014 for recent evidence that anticipatory coarticulation plays a role in spoken word recognition). Even if information from the speech signal affects auditory processing in a delay of exactly 100 ms, then, much of this information may precede the time in the word that was annotated as the beginning of the segment. We therefore explore a 0 ms lag as well.

Finally, Ettinger et al. (2014) explore lags of 100 ms, 150 ms and 200 ms, and find that segment surprisal was most strongly predictive of neural activity with a 200 ms lag. However, the effect was only evident after the end of the word, beginning around 550 ms into the stimulus (i.e., 750 ms into the neural recording); this suggests that the success of the 200 ms lag is particularly relevant to the word's offset.

In sum, we present results with lags of 0 ms, 100 ms and 200 ms, while keeping in mind that a 100 ms lag is the most plausible one.

We calculated surprisal estimates for each segment as outlined in section 5.2.1.1 above. The duration of each segment was read off the alignments between the phonetic transcriptions and the sound files. We added a special end-of-word segment [#], which lasted 100 ms in all words. We made the simplifying assumption that the surprisal of a segment affects neural activity throughout

the duration of the segment. If the lag is 100 ms, for example, the surprisal of a segment may serve as a predictor of neural activity in a window between 100–150 ms and 100–300 ms after its onset, depending on the duration of the segment.

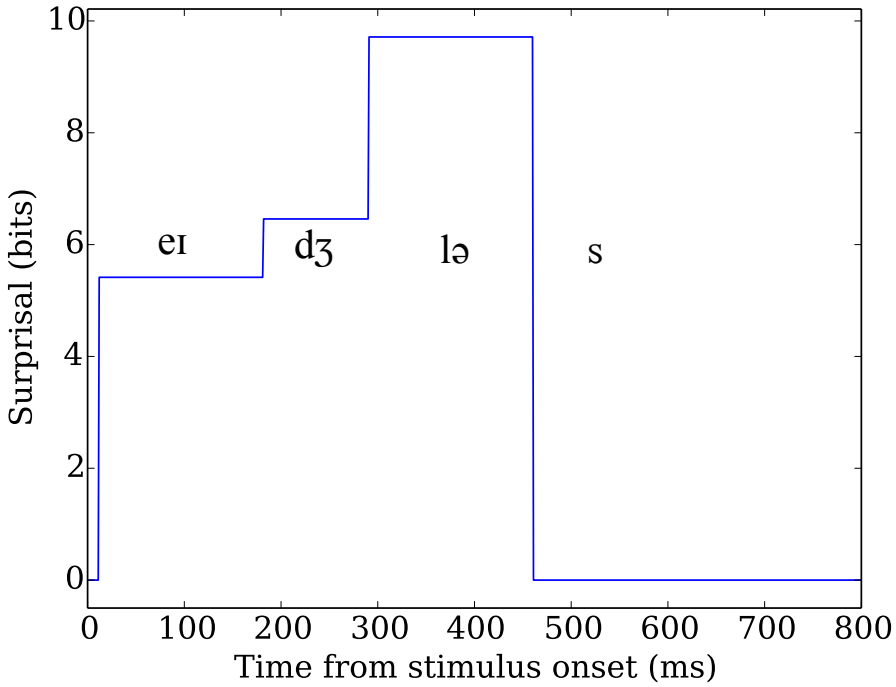


Figure 5.3: Timepoint-by-timepoint surprisal estimates for *ageless*. Note that the string [eidʒl] uniquely identifies the word in the corpus we used; the surprisal of the last segment [s] is therefore 0.

5.3 Results

5.3.1 Behavioral

Accuracy rates ranged between 83% and 98% (median 92%). Mean reaction times varied widely between 381 ms and 2488 ms (median 826 ms), reflecting the fact that the task was not speeded. Three of the paraphrases had accuracy rates below 83% (*rudeness*: 50%, *weakling*: 62%, *keenly*: 71%). An inspection of the sound files suggested that the particularly low accuracy rate for *rudeness* was due to the recording of the paraphrase sounding ambiguous between *in a polite way* (the

intended paraphrase) and *impolitely*. There were no obvious acoustic issues with the remaining two items. We decided not to exclude any items from the analysis of the neural data because we did not have accuracy data for most of our items. To test whether accuracy rates differed across conditions, a logistic mixed-effects regression model was fit to the responses to the paraphrases, excluding filler items and the three low-accuracy words mentioned above. Models were fit using the *lme4* package (Bates et al., 2014) in R, with a by-item and by-subject random intercepts and by-subject slopes for entropy and surprisal. Higher entropy was marginally associated with higher accuracy ($\beta = 0.67, p = 0.09$, likelihood ratio test); accuracy was not affected by surprisal and frequency.

5.3.2 Average activity

In sensor space, participants showed a pattern typical of auditory experiments, dominated by a component peaking around 100 ms after stimulus onset (M100; see Figure 5.4). Figure 5.5 shows the dSPMs averaged across subjects. To compensate for inter-subject variability, each subject's dSPM was normalized before averaging by dividing all of the values in the dSPM by the maximal value for that subject's dSPM. Early on activity is localized to areas around auditory cortex; as the word progresses, activity spreads across the insula and the superior temporal and inferior temporal gyri. A surprising amount of the activity localizes to the insula; this may reflect a localization error.

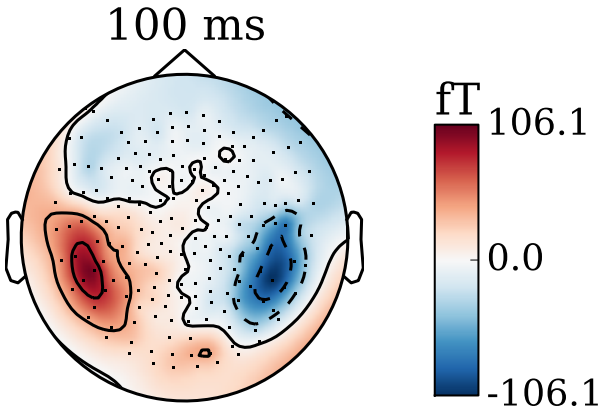
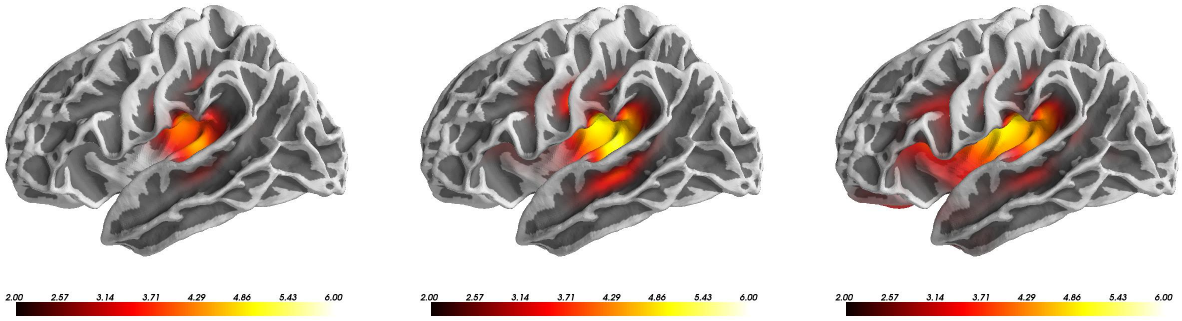


Figure 5.4: Sensor space activity 100 ms after the onset of the stimulus (a representative subject).



(a) (b) (c)

Figure 5.5: Grand average of normalized dSPMs: (a) 100 ms after stimulus onset; (b) at the morpheme boundary; (c) at the offset of the word.

5.3.3 Regions of interest

Regions of interest were defined based on the Desikan-Killiany parcellation included in FreeSurfer (Desikan et al., 2006); see Figure 5.6. We did not have structural MRIs for our subjects; the anatomical ROIs were therefore defined based on the average brain provided by FreeSurfer.

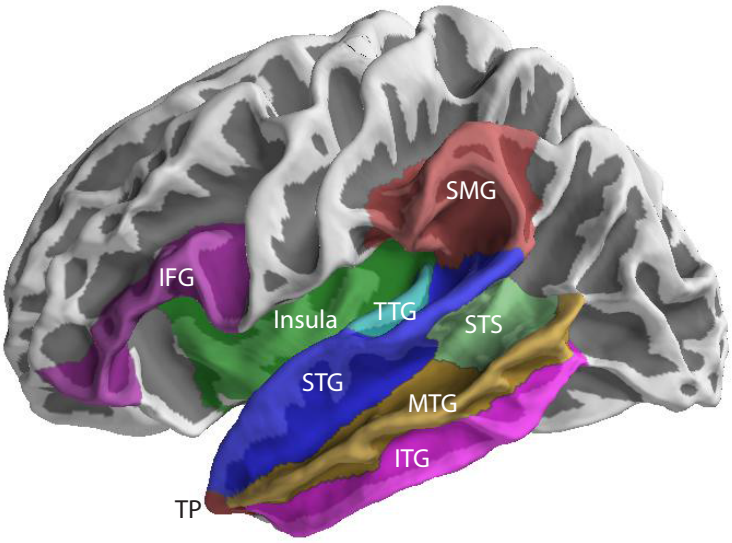


Figure 5.6: Language network. IFG: inferior frontal gyrus, ITG: inferior temporal gyurs, MTG: middle temporal gyrus, SMG: supramarginal gyrus, STG: superior temporal gyrus, STS: superior temporal sulcus, TP: temporal pole, TTG: transverse temporal gyrus.

The grand average (Figure 5.5) indicates that most of the signal was localized to the transverse temporal gyrus (TTG, which encompasses the auditory cortex), as well as the insula. Below we will consider the TTG as a primary region of interest. While the TTG is not itself a language area, prediction error often manifests in low-level perceptual regions (Dikker et al., 2009), in accordance to the predictive coding hypothesis. Ettinger et al. (2014) report that the clearest prediction error signal was registered in TTG.

Prediction error may also affect activity in the left superior temporal gyrus (STG), which is considered to the hub of speech perception (Scott & Johnsrude, 2003; Hickok & Poeppel, 2007) and is sensitive to the phonological features of the sound being perceived (Mesgarani, Cheung, Johnson, & Chang, 2014). Most pertinently, it shows predictability effects in spoken word recog-

nition (Ettinger et al., 2014; Gagnepain et al., 2012; Sohoglu et al., 2012). We note that little of the activity was localized to the STG, perhaps reflecting a localization error.

To detect activity related to the predictors of interest in higher-level language regions, we define a broad region of interest that includes left-hemisphere language regions (Hickok & Poeppel, 2007): the left inferior frontal gyrus (LIFG), the insula, the supramarginal gyrus and the inferolateral temporal lobe (including the superior temporal sulcus, STG, TTG, middle temporal gyrus, inferior temporal gyrus and temporal pole).

5.3.4 Morphological prediction

Regression models with entropy, surprisal, log word frequency and the interaction of entropy and surprisal were fitted, followed by group level spatiotemporal cluster tests (as described above), within four regions: left TTG, left insula, left STG and the full set of left-hemisphere language areas. Each analysis was performed six times: time-locked to the onset of the word, the morpheme boundary and the end of the word, and for each of those options, once with inclusive predictors and once with exclusive ones.

The only predictor that had an (uncorrected) significant effect in any of the analyses was the interaction between entropy and surprisal in the inclusive analysis in the TTG ($p = 0.035$; see Figure 5.7). The cluster was found between 430 ms and 490 ms after the onset of the stimulus, that is, around 110 ms after the median location of the morpheme boundary. A marginal interaction between surprisal and entropy in the same direction as the one found in the TTG was found in the insula (again in the inclusive analysis), though somewhat later (510–570 ms after onset; $p = 0.06$); a similar marginal interaction cluster in that region was found in the exclusive analysis (500–550 ms after onset; $p = 0.08$). Finally, there was a marginal effect of the interaction in the inclusive analysis of the full set of language regions when time-locked to the offset (50–130 ms after offset; $p = 0.1$); Figure 5.7g shows that this cluster localized to the middle temporal gyrus, and that the interaction had a positive sign.

To illustrate the pattern of interaction, we performed median splits within each of the two variables and plotted a bar graph of the means within each of the four combinations (with the caveat that this simplified visualization does not reflect the continuous interaction and may be misleading in some cases). The crossover interaction in the insula and TTG does not readily correspond to any obvious cognitive hypothesis: There is a predictability trend in the expected direction for low competition stems, but a reverse predictability trend for high competition stems. If only the predictability effect for low competition stems was genuine this pattern would correspond to the conservative prediction hypothesis. Note, however, that both of these effects are timelocked to the onset of the stimulus rather than to the morpheme boundary, which may have been expected. The marginal offset effect in the MTG is even harder to interpret: it appears to show that predictable suffixes elicit more neural activity than unpredictable ones, but more so in low-competition scenarios.

5.3.5 Phonological prediction

5.3.5.1 Timepoint-by-timepoint

We report two statistical analyses of the effect of segment surprisal on neural activity. The first is a timepoint-by-timepoint analysis (Ettinger et al., 2014). Like the morphological prediction analysis described above, this analysis uses a spatiotemporal cluster test; in contrast with that analysis, however, the variable used to predict neural activity varies from timepoint to timepoint, based on the properties of the segment heard e.g. 100 ms earlier.

TTG/insula: 100 ms lag No cluster in the TTG reached significance; the closest cluster to the significance threshold included the activity corresponding to the segments played 500–530 ms into the recording, i.e., the neural activity at 600–630 ms ($p = 0.1$). A cluster in the insula around the same time (480–570 ms into the recording) was statistically significant ($p = 0.01$; see Figure 5.8c-d).

TTG/insula: 0 ms lag There was a marginally significant cluster in the TTG (580–630 ms, $p = 0.08$); see Figure 5.8a-b. This cluster was preceded by two slightly weaker clusters (440–480 ms, $p = 0.15$; 500–540 ms, $p = 0.1$). This appears to be a sustained effect in the range 440–630 ms, though occasional dips below the significance threshold break the cluster of uncorrected significant activity into three smaller clusters. There was an analogous cluster in the insula that did not reach significance (550–630 ms, $p = 0.13$).

Other analyses There were no significant effects in the 200 ms lag in any of the regions, nor any significant effects in the STG or the broad language area analysis in any of the lags.

5.3.5.2 Segment-based analysis

The second analysis treats the activity in a region of interest that corresponds to a segment in a particular trial as a single data point. In other words, if the segment lasts between times t_1 and t_2 and the lag is l, the dependent variable is the average neural activity in the region between $t_1 + l$ and $t_2 + l$. This analysis is closely related to typical analyses of reading time corpora (e.g., Roark et al., 2009).

We investigate the effect of segment surprisal in the left TTG, STG and insula, with lags of 0 ms, 100 ms and 200 ms. A preliminary test showed that the timepoint at which the segment started was a strong predictor neural activity in all cases: Segments that started later into the word elicited more neural activity. This effect leveled off as the word progressed (i.e., there was generally a significant negative quadratic term). We therefore fitted linear mixed-effects models with segment onset time as a control predictor (both a linear and a quadratic term), as well as segment surprisal and the interaction between surprisal and onset time. We included a by-subject random intercept and random slopes for surprisal and for the interaction between surprisal and onset time, as well as a by-segment random intercept. To facilitate comparing the effect size across regions and time lags, the predictors and the dependent variables were all standardized (i.e., the mean was subtracted and the result was divided by the standard deviation).

We assessed the statistical significance of surprisal in two ways: By comparing the log-likelihood of a baseline model with only onset time (linear and quadratic terms) and a the log-likelihood of a model with those terms and surprisal and the interaction between surprisal and onset time; and by adding the two terms sequentially.

Lag of 0 ms The effect of onset time was strong across regions[3] (linear terms – TTG: $\beta = 0.11$; STG: $\beta = 0.14$; insula: $\beta = 0.15$; all ps < 0.001).

There was a marginal effect of surprisal in the TTG (combined: $p = 0.08$; sequential: $p = 0.23$ for surprisal, $p = 0.05$ for interaction) but not in the other regions (ps > 0.1).

Lag of 100 ms Onset time again had a strong effect across regions, though roughly half as strong as the effect with a 0 ms lag (linear terms: TTG: $\beta = 0.06$; STG: $\beta = 0.08$; insula: $\beta = 0.09$; all ps < 0.001).

The effect of surprisal was highly significant in the TTG, though smaller than the effect of onset time (combined: $p < 0.001$; sequential: surprisal: $\beta = 0.02$, $p < 0.001$; interaction: $\beta = 0.02$, $p = 0.02$). The same effect was significant though weaker in the insula (combined: $p = 0.03$; sequential: surprisal: $\beta = 0.01$, $p = 0.02$; interaction: $\beta = 0.01$, $p = 0.2$), and marginal in the STG (combined: $p = 0.05$; sequential: surprisal: $\beta = 0.01$, $p = 0.02$; interaction: $\beta = 0.005$, $p = 0.52$).

Lag of 200 ms The effect of onset time was much weaker in this lag. In the TTG, its size was $\beta = 0.003$ for the linear term and $\beta = -0.02$ for the quadratic term; removing these two effects significantly decreased the log-likelihood of the model, though much less than in the shorter lags ($p = 0.03$). The analogous effects in the STG and insula were much stronger, though again weaker than in shorter lags (linear terms: STG: $\beta = 0.03$; insula: $\beta = 0.04$; both ps < 0.001).

[3] Since the predictors and outcomes were standardized, the regression coefficient indicate by how many standard deviations the outcome would change if the predictor were changed by a single standard deviation.

The effect of surprisal did not reach significance in any of the regions (all combined analyses ps > 0.1).

Influence of word-end surprisal The analyses above included the special 100-ms long segment [#] which marks the end of the word. To assess its contribution to the results, we repeated the analyses after excluding this segment. The central effect of surprisal in the TTG at a 100 ms lag was somewhat weaker though still significant (combined: $p = 0.004$; sequential: surprisal: $\beta = 0.01$, $p = 0.005$; interaction: $\beta = 0.01$, $p = 0.08$). The effect was no longer significant in the STG (combined $p = 0.17$) but was not diminished in the insula (combined $p = 0.02$). In summary, while the special word-end segment contributed to the effect of surprisal, it was not its only cause.

5.4 Discussion

5.4.1 Morpheme prediction

The target items were bimorphemic words that varied along two dimensions: the surprisal of the affix and the entropy of the morphological cohort after the stem. We ran a series of analyses to assess whether either of these factors or their interaction affects neural activity in left-hemisphere language arareas.

The only results that reached or approached the uncorrected significance threshold were interaction effects that were difficult to interpret. While the analyses we performed are certainly not independent, some correction for multiple comparison would certainly need to be performed; however, since the lowest p-value was quite close to the uncorrected significant threshold of 0.05, it is unlikely that any of the effects would survive such a correction. Combined with the difficulty in interpreting the pattern of results, we conclude that there were no conclusive effects of morphological prediction in the analysis we used. There was no clear advantage to using the exclusive predictors; if anything, the inclusive predictors were slightly more effective. While the lack of clear effects in either analysis precludes drawing definite conclusions, the slight superiority of inclusive

predictors suggests that listeners cannot tell with certainty whether a word will be monomorphemic or bimorphemic.

5.4.2 Segment prediction

We quantified the predictability of each segment based on the word frequency distribution of English, and assessed how this factor affected neural activity, in particular in auditory regions. We performed two types of analysis: a continuous timepoint-by-timepoint analysis in which the surprisal of the segment played at time t was used to predict neural activity at time $t + l$ (where l is the lag); a segment-by-segment analysis, in which neural activity between time $t_1 + l$ and $t_2 + l$ was averaged for a segment that was played between t_1 and t_2, and a regression model was fitted to predict this average activity from the surprisal of the segment. The most plausible value for l is 100 ms; however, we also explores lags of 0 ms and 200 ms.

The timepoint-by-timepoint analysis yielded a significant positive effect of surprisal in the insula at the 100 ms lag (we suspect that activity localized to the insula may in fact originate in auditory cortex), though only late in the word. An apparently sustained effect in the TTG was found in the second half of the word with a 0 ms lag. The fact that the effect of surprisal was evident in both lags and only evident at the end of the word is somewhat worrying, as it may reflect the fact that surprisal late in the word is confounded with the length of the word: the surprisal after the offset of a short word is 0, whereas some of the longer words may have segments with nonzero surprisal at the same time point.

This concern is somewhat alleviated by the segment-by-segment analysis showed an effect of surprisal in the 100 ms lag; this effect was particularly strong in the TTG and weaker in the insula and STG. This analysis does not compare the same timepoint in longer and shorter words (neural activity past the last segment isn't included in the analysis). There are still important confounds that need to be examined (e.g., words with earlier vs. later uniqueness points); however, the results of this analysis appear promising, and the fact that a single number is associated with

each segment, compared to the massive number of data points associated with each segment in the previous analysis, facilitates using accurate regression models and controls.

In estimating the predictability of a sound, we made the following simplifying assumptions, common in work based on the cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987; Ettinger et al., 2014):

1. Segment perception is veridical: listeners do not maintain uncertainty about the identity of the segments they have perceived so far (see Connine, Blasko, & Hall, 1991 for evidence to the contrary).

2. The representation that is being predicted is an unanalyzed segment rather than a set of phonological features; if a listener expected the next segment to be [g], he or she will be equally surprised to hear a similar segment ([b]) as a dissimilar one ([r]).

We made several additional assumptions to generate the timepoint-by-timepoint phonological surprisal estimates:

1. The prediction error associated with a segment is constant throughout the segment. In reality, the majority of the information conveyed by the segment is likely to be processed shortly after the segment is perceived.

2. There is a constant lag between the onset of the segment and the point at which it can be compared against expectation in auditory cortex. In reality, the information in some segments may not be available immediately at the onset of the segment; voiceless stops, for example, are probably not identifiable until they are released (except for their coarticulatory effect on the preceding vowel).

3. The prediction error associated with a segment only affects neural activity while that segment is being perceived. In practice, it is likely that it could last into the next segment, especially for shorter segments such as stops.

All of these assumptions are likely to be false, and should be examined in future work. Some of the assumptions can be tested in the present data set (e.g., the grain size of phonological prediction).

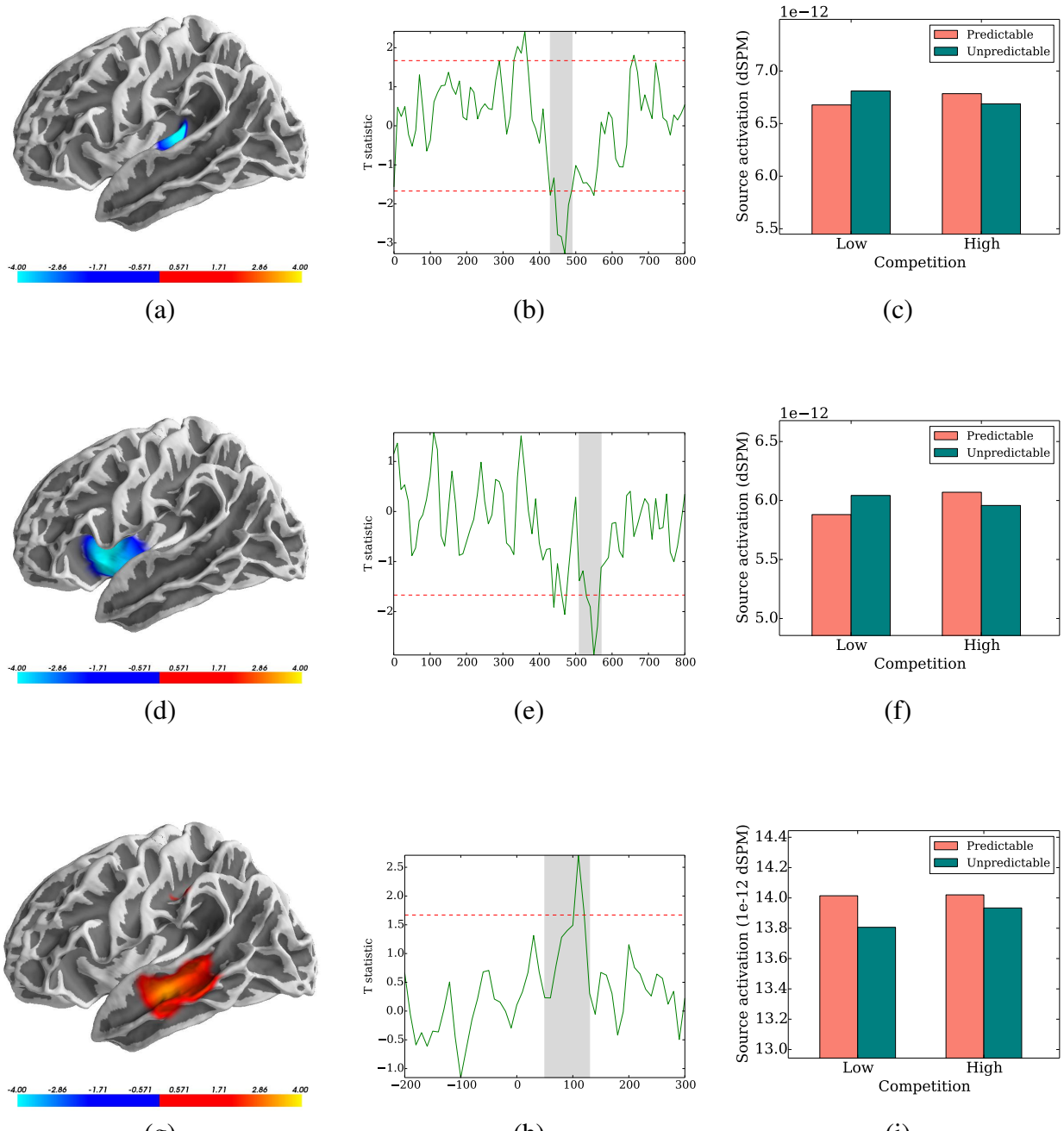


Figure 5.7: Morphological prediction: effects of interaction between (inclusive) surprisal and entropy. The leftmost figure in each row shows the spatial extent of the cluster; the middle figure shows the timepoint by timepoint average across subjects and sources in cluster of regression coefficient; and the right figure illustrates the direction of the interaction using a median split within both variables. The regions within which the spatiotemporal cluster test was run were: (a-c) TTG; (d-f) insula; (g-i) all language areas. In the TTG and insula the analysis was timelocked to the onset of the stimulus, whereas in the full language area analysis it was timelocked to its offset.

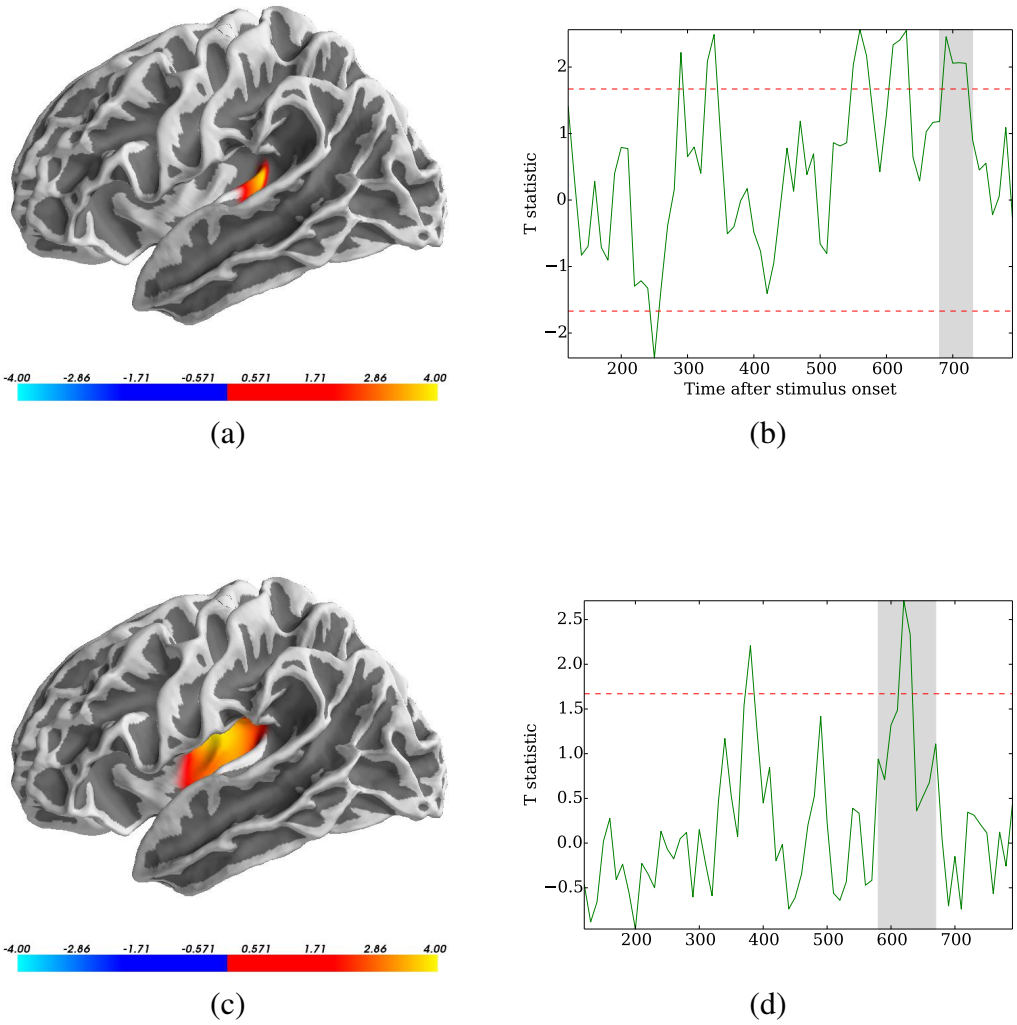


Figure 5.8: Timepoint-by-timepoint segment surprisal effects in the left TTG with a 0 ms lag (a-b) and in the insula with a 100 ms lag (c-d). (a, c) spatial extent of cluster; (b, d) Timepoint by timepoint group level t statistic for the surprisal regression coefficient averaged across the spatial extent of the cluster. The shaded area indicates a statistically significant cluster (note that since not all sources reach the threshold in all timepoints, the average t statistic may dip slightly below the threshold). The dashed horizontal lines indicates the threshold statistic for inclusion in a cluster ($|t| = 1.67$).

6 Rapid generalization in phonotactic learning

6.1 Introduction[1]

Natural languages often place restrictions on how sounds can combine to form words. The velar nasal [ŋ], for example, can occur in the coda of an English syllable, as in *ring* [ɹɪŋ] or *finger* [fɪŋgəɹ], but not in its onset:[2] Words like **ngir* [ŋɪɹ] do not exist in English (McMahon, 2002). English speakers do not typically consider this to be an accidental gap, and judge words that start with a [ŋ] as unlikely to become words of the language. The set of all such restrictions is referred to as the phonotactics of the language. The distinction between phonotactically legal and illegal words is reflected in a variety of implicit tasks, in both adults and infants (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; McQueen, 1998).

Sounds that share articulatory or perceptual features tend to have similar phonotactic distributions. German, for example, allows all voiced consonants (e.g., [b] or [g]) to occur anywhere in the

[1]Experiments 2a and 2b were published in the Proceedings of Phonology 2013 in a different form (Linzen & Gallagher, 2014). We thanks Frans Adriaans, Michael C. Frank, Timothy O'Donnell and Todd Gureckis for discussion, as well as audiences at the 8th Northeast Computational Phonology Circle and the 36th Annual Cognitive Science Society Meeting.

[2]The onset of a syllable consists of the consonant or consonants that precede its nucleus (the nucleus is generally a vowel); its coda consists of the consonant(s) that follow the nucleus. In the syllable [flɪp], for example, [fl] is the onset, [ɪ] is the nucleus and [p] is the coda. We use an asterisk to represent phonotactically illegal words.

word except its end: bal] is a valid German word, but *[*lab*] isn't. Although such patterns can in principle be captured by a disjunction of multiple sound-specific patterns, learners represent them as class-wide patterns as well (Saffran & Thiessen, 2003; Cristià & Seidl, 2008). For example, English speakers rate *srip* as a better potential word of English than *mbip* (Scholes, 1966; Daland et al., 2011), even though neither of the sequences [sɹ] and [mb] occur at the onset of English words. A plausible interpretation of this finding is that speakers generalize from the attested English onsets [sl] and [ʃɹ] (e.g., *slip* or *shrewd*) to the wider class of strident-liquid onsets, which also includes [sɹ]. On the other hand, there are no attested sonorant-stop onset sequences that would give rise to an analogous generalization that would apply to [mb] (Albright, 2009).

6.1.1 The time course of generalization

In line with these observations, models of phonotactic knowledge that incorporate natural-class based generalizations typically predict behavioral results better than models that do not (Adriaans & Kager, 2010; Albright, 2009; Berent, Wilson, Marcus, & Bemis, 2012; Hayes & Wilson, 2008). The relevant behavioral results are generally obtained by presenting novel strings to speakers of a given language and asking them for implicit or explicit judgments of how likely those words are to be words of that language; in other words, models are tested on their ability to simulate the end-state of language learning, after considerable exposure to the language. Conversely, there is little empirical data on the *time course* of phonotactic learning, even though those models make different assumptions about that time course. The goal of this paper is to characterize the process by which phonotactic generalizations are learned, with an eye to constraining models of phonotactics.

We contrast two views of the order of acquisition of item-specific knowledge and broader generalizations (see also Cristia & Peperkamp, 2012; Kapatsinski, 2014). In one view, which we will term *specific-to-general learning*, learners must first acquire knowledge about specific segments. Once they have noticed the commonalities among specific segments, they can form a generalization that may encompass unattested phonotactic structure in addition to the attested ones that gave rise to the generalization. For example, when acquiring the phonotactics of English,

learners must first learn that English syllables can start with [b] and that they can start with [g] before they can make the generalization that English syllables can start with a voiced stop (Albright & Hayes, 2003; Albright, 2009; Adriaans & Kager, 2010).

StaGe (Adriaans & Kager, 2010) is a particularly clear example of the specific-to-general view. It includes two distinct modules: a statistical learning module and a generalization module (for a related proposal, see Thiessen, Kronstein, & Hufnagle, 2013). The statistical learning module tracks the frequency of two-sound word-initial sequences (e.g., [bl] or [mp]).[3] If the frequency of the sequence is significantly lower than expected from the frequencies of the sounds that make up the sequence (typically less than half of the expected frequency), the model induces a constraint against having the sequence as a word-initial cluster. Whenever the statistical learning module has acquired constraints against two sound sequences that differ in exactly one phonological feature, the generalization module constructs a constraint that abstract away from that feature. For example, if the statistical learning module has acquired a constraint against [bl] and [gl], which differ only in the place of articulation of the first consonant, the generalization module will construct a constraint against all voiced stops followed by an [l] (which can then apply to [dl], for instance).

A second view, which we will term *simultaneous learning*, does not assume that the learning of a generalization presupposes the existence of sound-specific knowledge. The existence of a [b]-initial syllable in the input supports both a segment-specific pattern (e.g., syllables can start with [b]) and a more general one (syllables can start with a voiced stop, or even a stop in general). Consequently, learners may acquire class-wide patterns before any of their segment-specific instances. This view underlies Maximum Entropy models (Hayes & Wilson, 2008; Pater & Moreton, 2012) and PAIM (Linzen & O'Donnell, 2015).

[3] StaGe is a model of phonotactics-based word segmentation. We abstract away from the details related to word segmentation and focus on the phonotactic learning aspect of the model.

6.1.2 Overview of the paper

This paper reports the results of a series of artificial language learning experiments that probe the time course of phonotactic generalization. In Experiment 1, participants were taught a language in which all word onsets had the same value for the voicing feature (e.g., all were voiced). This knowledge could be represented either as a disjunction of segment-specific patterns (onsets can be [b], [d] or [g]) or as an abstract pattern (onsets are voiced). Participants were divided into several groups, each of which received a different amount of exposure to the language. After the exposure phase, participants judged novel test words for acceptability. If participants preferred test words whose onset appeared in exposure to test words with a new onset, that would constitute evidence for learning of the specific phonotactic patterns. Conversely, a preference for test words whose onset shared the voicing value with the exposure words to test words that didn't, regardless of whether the specific onset appeared in the exposure phase, would provide evidence for learning of the abstract pattern. Participants showed evidence of learning the abstract pattern before they showed evidence of learning the specific ones; the results of Experiment 1 were therefore inconsistent with the specific-to-general view.

The abstract regularity in Experiment 1 was a categorical phonotactic restriction based on a phonetic feature. Experiment 2a tested the generality of the findings by teaching participants a language with an abstract generalization that is not tied to a phonetic feature—specifically, identity between two consonants. Moreover, this regularity was probabilistic rather than categorical. The results were qualitatively similar to the results of Experiment 1. Experiment 2b taught participants a control language whose goal was to verify that the results of Experiment 2a were indeed due to learning rather than pre-existing biases.

Finally, Experiment 3 tested whether learners can acquire a general phonotactic regularity based on a single instance of the regularity. Participants indeed generalized to sounds that were not in the exposure set but shared a phonetic feature with the single attested exposure sound, indicating that generalization to an abstract pattern does not require extracting commonalities between multiple specific patterns.

6.2 Experiment 1: A natural-class based generalization

The artificial language used in this experiment was modeled after the one used by Cristia, Mielke, Daland, and Peperkamp (2013). It had a categorical natural-class based phonotactic generalization: All word onsets had the same voicing (either all voiced or all voiceless; different versions of the language were presented to different participants). Following the exposure phase, participants provided acceptability judgments on words of three types:

1. Conforming attested onset (CONF-ATT): words whose onset appeared as the onset of one of the exposure words. Since the phonotactic pattern was categorical, all such onsets conformed to the generalization.

2. Conforming novel onset (CONF-UNATT): words whose onset did not appear as the onset of any of the exposure words, but had the same voicing as those onsets.

3. Nonconforming unattested onset (NONCONF-UNATT): words whose onset differed in voicing from the onsets of all of the exposure words.

All of the test words were distinct from the exposure words. This was the case even for CONF-ATT test words, where the onset (C_1) was shared with some of the exposure words, but the full word ($C_1V_1C_2V_2$) was novel.

Exposure sets were constructed which consisted of five words, one with each of the exposure onsets. Participants were divided into four groups; each group was given a different number of exposure sets (one, two, four or eight). For example, participants in the One Set group heard five exposure words, one with each of the exposure onsets, and participants in the Two Sets group heard ten exposure words, two with each of the exposure onsets. Participants were not given any indication that the exposure words were organized into sets. A detailed description is given in the Materials section below; see Table 6.1 for examples.

The focus of the study is the relationship between the amount of exposure to the language that participants receive and the knowledge that they extract from that input. The specific-to-general

Exposure	Test		
kelo	CONF-ATT	CONF-UNATT	NONCONF-UNATT
tanu	falu	soma	zila
fula	femi	sunu	zoma
θomi			
pinu			

(a)

Exposure		Test		
ganu	gimi	CONF-ATT	CONF-UNATT	NONCONF-UNATT
balu	bini	zini	dimu	talu
vimu	voni	zonu	dila	tumu
zalu	zili			
ðano	ðamu			

(b)

Table 6.1: Two examples of the materials presented to participants in Experiment 1. (a) One exposure set, voiceless exposure onsets, [s] held out; (b) Two exposure sets, voiced exposure onsets, [d] held out.

ATT (attested): onset consonant (but not the full word) was encountered in exposure phase.
UNATT (unattested): onset consonant were not encountered in exposure phase.
CONF (conforming): onset consonant conforms to the abstract pattern.
NONCONF (nonconforming): onset consonant does not conform to the abstract pattern.

view would be most clearly supported if at first participants only distinguished CONF-ATT onsets from all other onsets after a small amount of exposure, and then began generalizing to CONF-UNATT onsets as they received more exposure. The simultaneous view would be most strongly supported if participants first acquired the voicing generalization and only then the specific onsets; that is, if they showed evidence of distinguishing CONF-ATT and CONF-UNATT onsets on the one hand from NONCONF-UNATT onsets on the other hand before they showed evidence of recognizing CONF-ATT onsets. Finally, an outcome where participants started making both of these distinctions at the same time would be compatible with both views.

6.2.1 Method

6.2.1.1 Materials and procedure

The materials were a simplified version of those used by Cristia et al. (2013), adapted to English-speaking participants.[4] The onsets of all of the stimuli used in the experiment were drawn from the set of six voiced obstruents [b], [d], [g], [ð], [v] and [z] or from the set of six voiceless onsets [p], [t], [k], [θ], [f] and [s]. Words of the format $C_1V_1C_2V_2$ were created with all possible combination of these onsets as C_1; the vowels [a], [e], [i], [o] and [u] as V_1; the consonants [l], [m] and [n] as C_2; and the vowels [a], [i] and [u] as V_2. When the resulting combination formed an existing English word, one of the consonants [l], [m] or [n] was added to the end of the word (e.g., tunal instead of tuna).

The words in the language, as in all others languages used in this paper, were stressed on their first syllable. The words were recorded by a native English speaker. The recordings were made at a sampling rate of 44.1 kHz in a sound-attenuated booth on a Marantz PMD-660 solid state recorder using a head-mounted Audio Technica ATM75 microphone.

Participants were assigned to one of 12 lists. All of the exposure words in each list had the same voicing: They were either all voiced or all voiceless. Five of the onsets were presented to the participants in exposure, and the sixth was held out. List 11, for instance, had exposure words with the onsets [p], [θ], [k], [f] and [t], but not [s]. Whether the exposure onsets were all voiced or all voiceless was counterbalanced across participants, as was the identity of the held-out onset. For each list, one of the onsets showed in exposure was selected as the onset for the CONF-ATT test condition (e.g., for List 11 the CONF-ATT consonant was [f] as in *fumi*). The CONF-UNATT onset was the onset from the same voicing class as the exposure that was held out (in List 11, [s] as in *sona*), and the onset of the NONCONF-UNATT words was the consonant with the opposite

[4]Specifically, the fricatives [ʒ] and [ʃ] were replaced with [ð] and [θ], respectively, since [ʒ] is rare in onset position in English. Furthermore, Cristia et al. (2013) had two types of nonconforming test onsets, "near" and "far"; our NONCONF-UNATT test condition only included their "far" onsets.

voicing to the CONF-UNATT one (in List 11, [z] as in *zili*). Tables 6.1a and 6.1b illustrate the full set of materials in the one exposure and two exposures group respectively, each with a different counterbalancing list.

The list of exposure words was constructed in blocks, such that each consecutive block of five words had exactly one word starting with each of the five exposure onsets. Participants did not receive any indication of the structure of the lists. The order of onsets was pseudo-randomized within each block. Likewise, the segments selected for the V_1, C_2 and V_2 slots were pseudo-randomized in consecutive blocks such that each block contained all possible segments for the relevant slot. The test words were presented in two blocks of three tokens, one token for each of the onsets representing the CONF-ATT, CONF-UNATT and NONCONF-UNATT categories, in pseudo-random order (again without indication of the division into two blocks).

6.2.2 Procedure

All experiments in this paper were conducted using Experigen, a JavaScript framework for running online experiments (Becker & Levine, 2010). Participants were recruited though Amazon Mechanical Turk (www.mturk.com), a crowdsourcing site that enables recruiting a large number of participants for a modest cost. Results obtained using Mechanical Turk have been repeatedly shown to replicate established findings from the experimental behavioral research literature (Crump, McDonnell, & Gureckis, 2013); this inspires confidence in the platform's adequacy as a source of participants for behavioral experiments. Participants were paid $0.65 for completing an experiment. They were told that they needed to be native speakers of English to complete the experiment. They were asked in a short demographic survey at the end of the experiment what their native language was; data from participants who reported a native language other than English were removed. Participants were limited to those with IP addresses within the United States. We rejected participants who performed multiple experiments or multiple versions of the same experiment, and assigned the task to new participants to reach the intended sample size.

The experiments were split into an exposure phase and a test phase. During the exposure phase, the participants listened to words from the artificial language. The words were presented in isolation—i.e., not in a continuous stream. Participants were told that the exposure phase would be followed by a test phase during which they will be required to decide if new words sound like they could belong to the language they were listening to (for a similar task, see Moreton, 2008, 2012; Reeder, Newport, & Aslin, 2013). During the test phase, the instructions for the task were repeated after every test word. Only two answers were possible: "yes" and "no".

6.2.2.1 Participants

Six participants completed each combination of the 12 lists and four exposure groups, for a total of 288 participants (72 participants per exposure group). Three participants were rejected because their reported native language was not English. We report data from the remaining 285 participants (116 women, 166 men, three unreported; median age: 30, age range: 18–68, one unreported).

6.2.2.2 Statistical analysis

Logistic mixed-effects models (LMEM) (Baayen, Davidson, & Bates, 2008; Jaeger, 2008) were fitted to the participants' responses ("yes" or "no") using version 1.1.6 of the lme4 package in R (Bates et al., 2014). To simplify the exposition, we present two separate sets of analyses, one comparing NONCONF-UNATT to CONF-UNATT onsets and another comparing CONF-UNATT to CONF-ATT onsets. We fitted two types of models: full models, which included all participants, and within-group models, which only included participants in a given group (e.g., the Two Sets group). Fixed effects in the full models included the group as a four-level factor and the onset type as a two-level factor. The random effect structure for all models included a by-subject intercept and a by-subject slope for the effect of onset type, as well as a by-onset intercept. Statistical significance was assessed using the log-likelihood ratio test: Terms were added sequentially to the model and the improvement in log-likelihood was assessed using the chi-squared distribution (this is the LMEM equivalent of a Type I ANOVA).

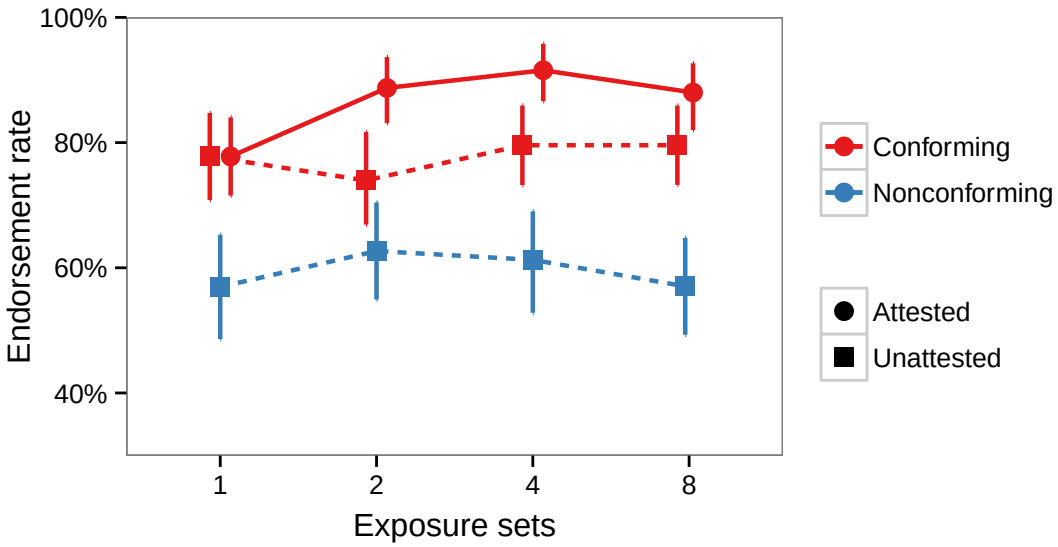


Figure 6.1: Mean endorsement rates for Experiment 1. Error bars represent bootstrapped 95% confidence intervals.

6.2.3 Results

The mean proportion of test words that participants in each group judged as acceptable in each of the conditions is shown in Figure 6.1. Endorsement rates were generally high: Even words that started with a NONCONF-UNATT onset were rated as acceptable around 60% of the time. This likely reflects the fact that apart from the onset these words matched the exposure words in all other respects; for example, word length, syllable structure, vowel pattern and identity of medial consonant were all consistent with the exposure words. Thus, all of the testing words received some support from the exposure data. We predict that words with novel syllable structures (e.g., *kes*) or longer words (e.g., *lumpumilu*) will be endorsed at a lower rate.

6.2.3.1 CONF-ATT vs. CONF-UNATT

In the full model, which included all participants, there was a significant main effect of onset type ($\chi^2(1) = 25.76$, $p < .001$), such that CONF-ATT onsets were rated more highly than CONF-UNATT ones. The main effect of group did not reach significance ($\chi^2(3) = 4.85$, $p = .18$), indicating that the average rating collapsing across these two onset types was similar across exposure groups.

The interaction between onset type and group was significant ($\chi^2(3) = 8.79$, $p = .03$). Within-group models showed that this interaction was driven by the absence of a significant preference for CONF-ATT onsets in the One Set group ($\chi^2(1) = .95$, $p = .33$), compared with a significant preference for CONF-ATT onsets in the Two and Eight Sets groups and a marginal preference in the Four Sets group (Two Sets: $\chi^2(1) = 8.1$, $p = .004$; Four Sets: $\chi^2(1) = 3.15$, $p = .08$; Eight Sets: $\chi^2(1) = 17.8$, $p < .001$).

6.2.3.2 NONCONF-UNATT VS. CONF-UNATT

CONF-UNATT words were rated more highly than NONCONF-UNATT ones ($\chi^2(1) = 32.27$, $p < .001$). The main effect of group was not significant ($\chi^2(3) = 0.57$, $p = .9$) and neither was the interaction between group and type ($\chi^2(3) = 2.82$, $p = .42$). Within-group models showed that the effect reached significance for all groups except for the Two Sets group, where the effect was in the same direction as in the rest of the groups but was only marginally significant (One Set: $\chi^2(1) = 12$, $p < .001$; Two Sets: $\chi^2(1) = 2.7$, $p = .1$; Four Sets: $\chi^2(1) = 8.08$, $p = .004$; Eight Sets: $\chi^2(1) = 11.2$, $p < .001$). Since the interaction was not significant, we do not interpret the difference between the Two Sets group and the other groups any further.

6.2.4 Discussion

Participants in Experiment 1 were taught artificial languages that had a categorical natural-class based phonotactic regularity: All word onsets had the same voicing (either all voiced or all voiceless, depending on the list). Participants then judged the acceptability of novel words with onsets of three types: CONF-ATT onsets, which were encountered during exposure; CONF-UNATT onsets, which shared the value for the voicing feature with the onsets of the exposure words but were not encountered during exposure; and NONCONF-UNATT onsets, which had the opposite value for the voicing feature than the exposure words. CONF-UNATT onsets were consistently endorsed more often than NONCONF-UNATT onsets, regardless of the amount of exposure: Even after a single set of exposure to each onset type, participants preferred onsets with the same voicing as the onsets of

exposure words to onsets with the opposite voicing. Conversely, participants did not start distinguishing CONF-ATT from CONF-UNATT onsets until after two or more exposures. This pattern of results favors the simultaneous view over the specific-to-general view: Participants in the One Set group, who did not show evidence of learning any of the individual onsets, still preferred onsets conforming to the generalization to onsets that did not.

The three-way distinction between CONF-ATT, CONF-UNATT and NONCONF-UNATT words was similar in the Two Set, Four Set and Eight Set groups. Despite growing evidence that not all generalization-conforming onsets are possible in the language, then, participants continued to generalize beyond the attested onsets.

6.3 Experiment 2a: A probabilistic abstract generalization

Participants in Experiment 1 showed evidence of learning a broad regularity (voiced stops can be onsets) before showing evidence of learning narrow regularities (e.g., [b] can be an onset). The broad regularity in Experiment 1 had two properties that may limit the generality of the conclusions that can be drawn from these results.

First, the generalization in Experiment 1 was categorical: It held of all of the words in the language. There is evidence that speakers' knowledge of the distribution of sounds in their language is not limited to the categorical distinction between possible and impossible: They also keep track of the relative frequencies of the possible sounds and sound sequences. Neither of the nonwords *riss* [ɪɪs] and *yowdge* [jaʊdʒ], for example, contains any sounds or sound sequences that are unattested in English; yet *riss*, which is comprised of frequent sound sequences, is judged to be a more likely potential word of English than *yowdge* (Coleman & Pierrehumbert, 1997).

Second, the generalization in Experiment 1 was stated over a phonetically defined class of sounds. While many phonotactic generalizations in natural language are based on the phonetic properties of individual sounds (e.g., "voiced stops are illegal codas" in German), some generalizations involve *relations* across sounds. A prominent example is the constraint against identical

root-initial consonants in Semitic languages (Greenberg, 1950; McCarthy, 1986; Frisch & Zawaydeh, 2001; Rose & King, 2007). In Hebrew, as in other Semitic languages, typical roots consist of three consonants (e.g., *ktb* 'write'). While roots with two identical consonants are in general fairly common (*smm* 'drug', *zrz* 'hasten'), cases in which the two identical consonants are the first two consonants of the root are rare to nonexistent (**ssm, *zzr*). Hebrew speakers are sensitive to this restriction. In a lexical decision task, for example, novel roots with initial duplication (e.g., *ssk*) are identified as nonwords faster than are novel roots with final duplication (*kss*); the preference for *kss* over *ssk* is stronger than the phonotactic probabilities of the sound sequences that make up those roots would predict (Berent, Shimron, & Vaknin, 2001). More strikingly, root-initial duplicated consonants are disfavored even when they are made up of consonants that do not exist in Hebrew. Although [w] is not part of the sound inventory of Hebrew, and consequently the sequence *ww* has a phonotactic probability of 0 both root-finally and root-initially, Hebrew speakers recognize **wwp* as a nonword faster than they recognize **pww* (Berent, Marcus, Shimron, & Gafos, 2002). This suggests that Hebrew speakers generalize over individual sequences of duplicated consonants to form a constraint disfavoring any root-initial duplicated consonant: *XXY, for any X and Y. Relational phonotactic generalizations have been documented in such diverse languages as Yucatec Mayan, Muna (an Austonesian language) and Peruvian Aymara (for a recent review, see Gallagher, 2013).

To replicate the findings of Experiment 1 and broaden the scope of the conclusions that can be drawn from those findings, Experiment 2a tested whether the pattern of results held for a probabilistic abstract generalization. All of the words in the language used in this experiment had the form $C_1V_1C_2V_2$ (e.g., *semi*). Vowels in the language varied freely, and the consonant pairs followed one of eight narrow phonotactic regularities. Four of those regularities involved two different consonants, e.g., C_1 = [k] and C_2 = [s] (two words conforming to this particular regularity are *kesa* and *kisu*); the other four involved two identical consonants, e.g., C_1 = [p] and C_2 = [p] (as in *pepu*), or C_1 = [s] and C_2 = [s] (as in *sase*).

While the phonotactics of the language can be captured precisely using these eight narrow regularities, it was also the case that half of the words in the language conformed to the abstract regularity $C_1 = C_2$, much more than would be expected by chance. If participants learned this abstract generalization, they should generalize it to words that contain identical consonants outside of those included in the exposure phase.

As in Experiment 1, exposure sets were created that included exactly one word that conformed to each of the narrow regularities, for a total of eight words per exposure set (see Table 6.2). The language was taught to several groups of participants, each receiving a different number of exposure sets. In the test phase, participants were presented with new words that had either consonant pairs that were familiar from the exposure phase (ATT) or new consonant pairs (UNATT), and asked them to judge whether the testing words could belong to the language they had learned. Half of the new consonant pairs in testing had identical consonants (CONF) and half had non-identical consonants (NONCONF). By contrast with Experiment 1, it was possible to construct ATTESTED-NONCONFORMING test words, since the generalization was probabilistic and held only of half of the exposure words. This led a fully crossed design that allowed us to test for the independent contribution of the broad and narrow regularities.

If participants learned the broad regularity, namely that identical consonant pairs are particularly common in the language, they should prefer words with identical consonant pairs to words with non-identical consonants. If participants learned the narrow consonant-specific regularities, they should prefer words with attested consonant pairs to words with unattested ones.

6.3.1 Method

6.3.1.1 Materials and procedure

All words in the experiment were of the form $C_1V_1C_2V_2$, e.g., *kesa*. The exposure words had one of eight different consonant pairs, four of which were identical and four of which were not (see Table 6.2). All participants were presented with 16 testing words, eight with the consonant pairs heard in exposure and eight with new consonant pairs. Each of the individual consonants C_1 and

Exposure	Test	
CONF	CONF-ATT	CONF-UNATT
pipa	papi	keku
ʃuʃe	ʃeʃu	sasi
gapu	gugi	dʒidʒe
nuni	ninu	mamu
NONCONF	NONCONF-ATT	NONCONF-UNATT
kesa	kasi	pina
mudʒe	medʒa	nage
dʒuke	dʒuke	gaʃe
sami	sami	ʃipu

Table 6.2: Materials presented to the participants in Experiment 2a. The table shows a complete exposure and test set for the One Set group.

C_2 in the new consonant pairs were encountered during the exposure phase, in both initial and medial position, but not as a combination. A total of 12 unique words were constructed for each consonant pair, by crossing the pair with all non-identical combinations of [a e i u] in V_1 and V_2; e.g., for [p p], the words constructed were *pipa*, *pipe*, *pupa* and so on. The stimuli were recorded by a female native English speaker.

In the exposure phase, participants listened to one, two, four or eight exposure sets. All exposure words differed from each other; that is, the same consonant pair was never heard with the same vowels more than once. There were 16 test words in all groups, one item with each consonant pair. As in Experiment 1, the specific words from exposure phase were never repeated in the test phase. For example, if *bagu* and *biga* appeared in the exposure phase, neither could appear in the test phase, but *bega* could. Items were pseudo-randomized in blocks as in Experiment 1.

6.3.1.2 Participants

A total of 280 participants completed the experiment, 70 in each group. Demographic information was not collected due to a technical failure.

6.3.1.3 Statistical analysis

As in Experiment 1, we fitted a full model that included participants from all four groups, as well as within-group models for each of the groups. The full model had three fixed effects: one between subjects (the exposure group) and two within subjects (Attestation and Conformity). The random effect structure for subjects in the full model included an intercept and random slopes for Attestation, Conformity and the interaction between the Attestation and Conformity. We were unable to include a by-item random effect due to model convergence issues. As before, p-values were calculated using the chi-square approximation to likelihood ratio tests in a stepwise regression. In the within-group models as well it was necessary to simplify the random effect structure. All inferences involved models with random intercepts for subjects and for consonant pair. For inferences involving Attestation, we only included a random slope for the factor; likewise, for inferences involving Conformity, we only included a random slope for Conformity. The interaction between Attestation and Nonconformity was assessed in the model that only had an Attestation random slope, again for convergence issues. Inferences without a random slope can be anti-conservative (Barr, Levy, Scheepers, & Tily, 2013). As such, we can trust the model when it finds that an interaction is nonsignificant, but cannot necessarily trust it when it finds the interaction to be significant.

6.3.2 Results

6.3.2.1 Full model

Figure 6.2 illustrates the mean endorsement rates for each group and condition. The full statistical model yielded a main effect of group ($\chi^2(3) = 45.12$, $p < .001$), reflecting the fact that endorsement rates were higher for participants who received more exposure to the language. There was also a main effect of Attestation, reflecting higher average endorsement rates for words with ATT than for words with UNATT consonants ($\chi^2(1) = 34.04$, $p < .001$), and a main effect of Conformity, reflecting higher average endorsement rates for CONF than for NONCONF words ($\chi^2(1) = 35.46$, $p < .001$).

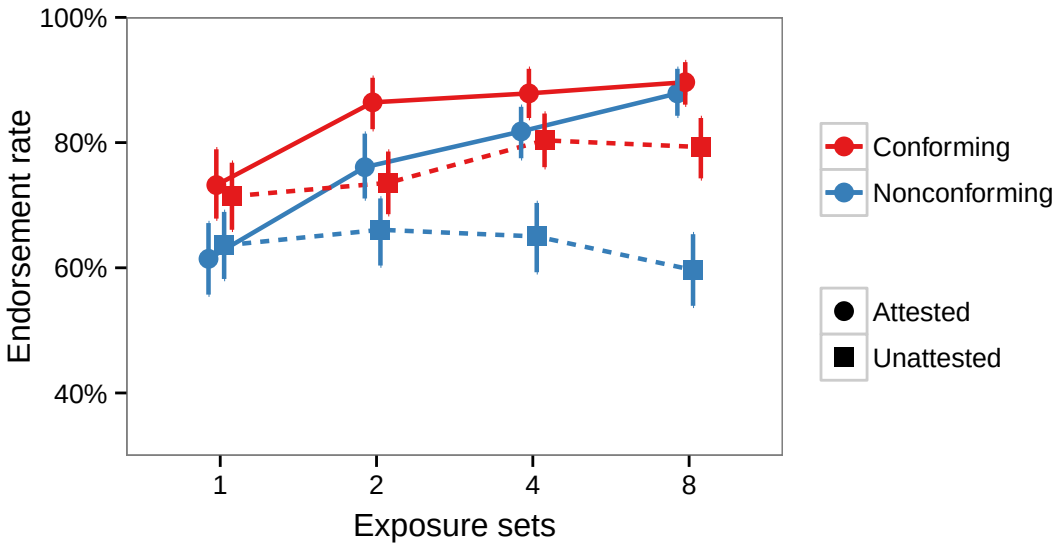


Figure 6.2: Mean endorsement rates for Experiment 2a. Error bars represent bootstrapped 95% confidence intervals.

The main effect of Attestation was modulated by an interaction with group ($\chi^2(3) = 37.9$, $p < .001$), which reflects the fact that participants were better at distinguishing ATT from UNATT items the more exposure they received to the language. The interaction of group and Conformity was not significant ($\chi^2(3) = 1.12$, $p = .77$), and neither was the interaction between Conformity and Attestation ($\chi^2(1) = 0$, $p = 1$). The interpretation of these findings is complicated by the significant three-way interaction ($\chi^2(3) = 8.59$, $p = .03$); Figure 6.2 suggests that the three-way interaction reflects the fact that as participants received additional exposure sets the effect of Conformity gradually diminished, but only for test words with ATT consonants; the effect of Conformity remained robust for test words with UNATT consonants even in the Eight Sets group.

6.3.2.2 Within-group models

One Set: In this group, CONF test words were rated significantly higher than NONCONF ones ($\chi^2(1) = 5.76$, $p = .02$). The main effect of Attestation and the interaction did not reach significance (Attestation: $\chi^2(1) = 0$, $p = .98$; interaction: $\chi^2(1) = .28$, $p = .6$), suggesting that the narrow phonotactic regularities did not affect endorsement rates. There was no sign of a nonsignif-

icant numerical trend towards an effect of Attestation (endorsement rates were: CONF-ATT: 73%; CONF-UNATT: 71%; NONCONF-ATT: 61%; NONCONF-UNATT: 64%).

Two Sets: Test words with ATT consonants were rated as acceptable significantly more often than ones with UNATT consonants ($\chi^2(1) = 12.1$, $p < .001$). The effect of Conformity was marginally significant ($\chi^2(1) = 3.56$, $p = .06$), and the interaction was nonsignificant ($\chi^2(1) = .99$, $p = .32$).

Four Sets: Both of the main effects reached significance; the interaction was again nonsignificant (Conformity: $\chi^2(1) = 4.97$, $p = .03$; Attestation: $\chi^2(1) = 11.49$, $p < .001$).

Eight Sets: The effect of Attestation was highly significant ($\chi^2(1) = 20.23$, $p < .001$). Conformity had a marginally significant effect ($\chi^2(1) = 3.49$, $p = .06$); however, there was a significant interaction between Attestation and Conformity ($\chi^2(1) = 4.23$, $p = .04$). As mentioned above, the absence of a random slope for the interaction makes the interpretation of this p-value problematic. However, separate models fitted within ATT and UNATT items (both with random found that Conformity had a highly significant effect for UNATT items ($\chi^2(1) = 17.6$, $p < .001$) but no discernible effect for ATT ones ($\chi^2(1) = 0.34$, $p = .56$).

6.3.3 Discussion

Participants showed evidence of learning the broad regularity over identical consonant pairs before they showed evidence of learning the narrower regularities over individual C_1–C_2 pairs. After a single exposure to each of the eight possible consonant pairs, four of which were pairs of identical consonants, participants showed a preference for novel words with identical consonants. This preference held regardless of whether or not this pair of identical consonants was presented in the exposure phase. Participants did not start showing evidence of learning individual consonant pairs until they received at least two sets of exposure. As in Experiment 1, this pattern of results suggests that broad regularities can be learned before narrower instances of those regularities.

Also echoing Experiment 1, participants consistently generalized to CONF-UNATT words even after eight exposure sets. To further explore this sustained generalization pattern, we administered the experiment to an additional group of 70 participants, this time with 16 exposure sets. Since we only had 12 distinct words with each consonant pair, some of the exposure words were repeated twice; it was still the case, however, that none of the test words occurred in the exposure phase.

The endorsement rates for the 16 Sets group were similar to the ones for the Eight Sets group, with the exception that the endorsement rate for NONCONF-UNATT words was more similar to the endorsement rate for those words in the other groups (One, Two and Four Sets); this suggests that the dip in endorsement rates for NONCONF-UNATT in the Eight Sets group visible in Figure 6.2 was spurious (CONF-ATT: 92%; CONF-UNATT: 79%; NONCONF-ATT: 89%; NONCONF-UNATT: 67%). Only the main effect of Attestetedness was significant ($\chi^2(1) = 28.72$, $p < .001$); the main effect of Conformity and the interactions were not (Conformity: $\chi^2(1) = 1.13$, $p = .29$; interaction: $\chi^2(1) = 1.23$, $p = .27$). The simple effect of Conformity was significant within UNATT words ($\chi^2(1) = 4.77$, $p = 0.03$) but not within ATT ones ($\chi^2(1) = .05$, $p = .83$). In sum, statistical evidence for generalization to CONF-UNATT words remained even for participants who received 16 exposure sets; the fact that this evidence was weaker than in the Eight Sets group may be to be an artifact of spuriously low endorsement rates for NONCONF-UNATT words in the Eight Sets group. In conclusion, participants continued applying the generalization to unattested consonant pairs even after ample evidence that only certain consonant pairs can appear in the language.

6.4 Experiment 2b: Ruling out an identity bias

We interpreted our participants' preference for identical items after one exposure set in Experiment 2a (the One Set group) as reflecting the learning of a probabilistic generalization that held of the exposure words. Before being confident in this interpretation, however, we must rule out the possibility that the preference for identical test items was due to prior bias favoring words with identical consonants rather than to exposure to the artificial language. Such a prior preference could

be derived from the participants' native language or from any number of perceptual or cognitive biases.

Experiment 2b was designed to rule out a pre-existing preference for words with identical consonants. Participants were exposed to words containing eight consonant pairs, all of which were non-identical. Since the question of interest relates to pre-existing bias, the only relevant exposure group is the One Set one. After the exposure phase, participants provided acceptability judgments on the same unattested items as in Experiment 2a (both CONF-UNATT and NONCONF-UNATT). If participants still showed a preference for identical over non-identical items, despite not having seen any identical items in exposure, this would be evidence that the preference is due to prior bias in favor of identical items. We refer to this hypothesis as the bias hypothesis. If, on the other hand, participants showed no identity preference, the interpretation of the identity preference in Experiment 2a as being due to learning would stand; we refer to this as the learning hypothesis.

6.4.1 Method

6.4.1.1 Materials and procedure

All words had the form $C_1V_1C_2V_2$, as in Experiment 2a. As in the One Set group of Experiment 2a, there were eight exposure words and 16 test words. All exposure words had two non-identical consonants (see Table 6.3). Vowel patterns were chosen at random, with no vowel pattern repeated across exposure and testing words. As in Experiment 2a, half of the test words were attested in exposure and half were not. All of the attested words in testing had non-identical consonants. The unattested words in testing had the same consonant pairs as in Experiment 2a, half identical and half non-identical (four of each). For consistency with Experiment 2a, we still use the labels CONF and NONCONF to refer to the test words with identical and non-identical consonants respectively, even though the exposure phase in Experiment 2b did not provide any evidence for the segment-identity generalization. Since no exposure words had identical consonants, there were no CONF-ATT test items; the three test conditions were NONCONF-ATT, CONF-UNATT and NONCONF-UNATT.

Exposure	Test	
	NONCONF-ATT	CONF-UNATT
ʃidʒa	ʃadʒi	keku
mune	mene	sasi
sagu	sugi	dʒidʒe
pusi	pisu	mamu
geka	gaki	
kupe	kepa	NONCONF-UNATT
nuʃe	nuʃe	ʃipu
dʒami	dʒami	pina
		nage
		gaʃe

Table 6.3: All consonant pairs used in exposure and test for Experiment 2b, with randomly selected example words.

The support that CONF and NONCONF test words received from irrelevant natural-class based patterns in the exposure set was matched as follows. Each of the eight consonants in the language appeared in the exposure phase once in initial position and once in medial position. As such, the CONF-UNATT and NONCONF-UNATT test words received equal support from the positional frequency of the individual consonants, as in Experiment 2a. In addition, CONF-UNATT and NONCONF-UNATT test words were matched for the amount of natural class based support they received from consonant cooccurences in the exposure word (voicing, place of articulation and manner of articulation). For example, the test word with the consonants [s]–[s] receives support from two voiceless-voiceless pairs ([p]–[s] and [k]–[p]), and there are no fricative–fricative pairs or alveolar–alveolar pairs in the exposure set, so its total natural class-based cooccurence support score is 2. It is matched with [g]–[ʃ], which also receives natural-class based support from two attested pairs, the single stop–fricative pair [p]–[s] and the single voiced–voiceless pair [g]–[k]; there are no velar-palatal pairs in the exposure set.

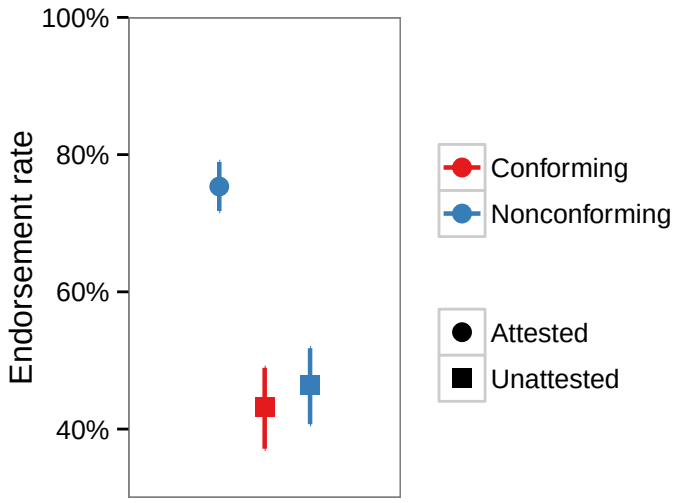


Figure 6.3: Mean endorsement rates for Experiment 2b. Error bars represent bootstrapped 95% confidence intervals.

6.4.1.2 Participants

A total of 70 participants completed the experiment (34 women, 35 men, one unreported; median age: 27, age range: 18-61).

6.4.1.3 Statistical analysis

A LMEM was fitted to the results, with a three-level factor of consonant type (NONCONF-ATT, CONF-ATT, CONF-UNATT) as a fixed effect, as well as random intercepts for consonant pair and for subject and a random slope by subject for consonant type.

6.4.2 Results

The results of Experiment 2b are shown in Figure 6.3. Contrary to the predictions of the bias hypothesis, participants did not show a preference for CONF-UNATT words; if anything, there was a slight preference for NONCONF-UNATT words over CONF-UNATT ones. There was a striking difference between NONCONF-ATT words and both CONF-UNATT and NONCONF-UNATT words:

Unlike the One Set group of Experiment 2a, participants in Experiment 2b were much more likely to endorse test words with attested than unattested consonant pairs.

Statistical analysis showed that the effect of condition on endorsement rates was highly significant ($\chi^2(2) = 27.6$, $p < .001$). We performed planned comparisons to examine the difference between the different levels of the factor. In line with Figure 6.3, the difference between NONCONF-ATT on the one hand and CONF-UNATT and NONCONF-UNATT on the one hand (i.e., the two UNATT conditions collapsed together) was highly significant ($\chi^2(1) = 27.4$, $p < .001$). By contrast, the difference between CONF-UNATT and NONCONF-UNATT did not approach significance ($\chi^2(1) = 0.57$, $p = .45$).

6.4.3 Discussion

Participants in Experiment 2b, who were not exposed to identical consonant pairs, did not show any preference for novel items with identical consonants (CONF-UNATT). The results therefore support the learning hypothesis, according to which the preference for identical items after one exposure in Experiment 2a was due to learning during the experiment. Thus, the interpretation of the main result of Experiment 2a remains unchanged: Participants showed evidence of learning the broad generalization about identical consonant pairs before learning narrow generalizations about the specific attested consonant pairs.

The results of Experiment 2b reveal an additional effect. Unlike in Experiment 2a, participants in Experiment 2b showed a strong preference for attested over unattested consonant pairs after just one exposure. While we cannot make firm claims about the source of this difference, one possibility is that the presence of a broad generalization interferes with the learning of narrower generalizations. In Experiment 2a, the presence of the identity generalization prevented learners from attending sufficiently to the narrower generalizations with small amounts of exposure, while in Experiment 2b learners were free to focus on the specific, attested consonant pairs.

At first blush, the lack of a preference for identical items in Experiment 2b compared to Experiment 2a could still be consistent with a pre-existing bias to give "yes" responses to identical

items: The absence of identical consonant pairs from the exposure data could have been taken as evidence for the generalization that pairs of identical consonants are underattested, offsetting a pre-existing bias in favor of identical consonants. However, this alternative explanation for the results of Experiment 2b becomes less plausible if we consider the radically different amount of support for the generalization that the exposure data provide in each of the experiments. With an inventory of 8 consonants, a sample of 8 words with all non-identical pairs is not a particularly surprising one: 56 out of the possible 64 consonant pairs are non-identical. The expected number of non-identical pairs in a sample of 8 is therefore 7, and an observed sample of 8 non-identical items yields an observed-over-expected ratio (O/E) of $8/7$. In Experiment 2a, on the other hand, the participants received four identical pairs instead of the expected one pair, for an O/E of $4/1$. In other words, the evidence for the overattestation of identical pairs in Experiment 2a is much stronger than the evidence for their underattestation in Experiment 2b. It is therefore implausible to assume that the preference for identical items after one exposure in Experiment 2a was due to bias, and at the same time that the lack of preference for identical items in Experiment 2b was due to learning that offset that bias.

6.5 Experiment 3: Generalization from a single type

Participants in Experiments 1 and 2a showed evidence of learning an abstract phonotactic generalization before they showed evidence of learning narrower, segment-specific ones. This argues against the specific-to-general approach to generalization: Learners do not need to first learn two or more special cases of a generalization before they can abstract away from these instances and form a generalization. Alternative explanations are still open, however. Participants in the One Set group of Experiments 1 and 2a may have learned only some of the specific types that support the generalization (e.g., two of the four CONF consonant pairs in Experiment 1), enough to form a robust bottom-up generalization, but not enough to produce a statistically significant difference between the endorsement rates for ATT and UNATT items. Alternatively, it may be the case that

participants are not confident enough in their knowledge of specific items to distinguish attested from unattested items when that knowledge is based on a small number of tokens, but still use that knowledge to form bottom-up generalizations (Albright & Hayes, 2003).

Experiment 3 rules out these alternative hypotheses by teaching participants a language in which only a single type supports a generalization. If participants still learn the generalization and apply it to types they did not see in exposure, this will provide additional support to the claim that generalizations do not have to be formed in a bottom-up fashion. Specifically, the exposure set contained only one type of voiceless stop onset (e.g., [p]); participants were tested to see if they endorsed the voiceless stops they had not encountered in the exposure phase (for example, [k] and [t], if [p] was the voiceless stop encountered in the exposure phase). Only two words starting with the voiceless stop were presented in exposure. Six filler words starting with onsets that were neither voiceless nor voiced stops were added to make the learning task more challenging and the exposure period longer. As in Experiment 1, participants rated three kinds of test items: CONF-ATT, CONF-UNATT and NONCONF-UNATT. We refer to this language as the Single Type language.

The experiment included two additional languages designed to allow us to draw firmer conclusions from the findings related to the Single Type language. The Two Types language included two different voiceless stops in the exposure set, e.g., [t] and [k]. One token was presented of each onset. Based on the results of Experiment 1, we expect participants assigned to the Two Types language to acquire the generalization, but not to distinguish attested from unattested onset types. Finally, the Control language did not include any voiceless stops at all: Participants who were assigned this language were only exposed to six filler words. This language served to examine whether participants had a pre-existing bias for or against voiceless stop onsets.

Exposure		Test		
FILLER	CONF	CONF-ATT	CONF-UNATT	NONCONF-UNATT
wamu	kami	kuna	pami	ðima
yuna	kamu		tanu	zanu
lani				
wina				
yani				
lima				

(a)

Exposure		Test		
FILLER	CONF	CONF-ATT	CONF-UNATT	NONCONF-UNATT
wamu	kami	kuna	pami	ðima
yuna	tamu	tanu		zanu
lani				
wina				
yani				
lima				

(b)

Exposure	Test	
FILLER	NONCONF-UNATT	CONF-UNATT
wamu	ðima	kuna
yuna	zanu	pami
lani		tanu
wina		
yani		
lima		

(c)

Table 6.4: Example of materials used in Experiment 3. (a) Single Type language, in the list that had [k] as the exposure CONF onset; (b) Two Type language, in the list the had [k] and [t] as the exposure CONF onsets; (c) Control language.

6.5.1 Method

6.5.1.1 Materials and procedure

Words were created with three classes of onsets: voiceless stops ([p], [t] and [k]), which we refer to as CONF onsets; voiced fricatives ([z] and [ð]), which we refer to as NONCONF onsets; and approximants ([w], [y] and [l]), which we refer to as FILLER onsets. All onsets were embedded in words of the form $C_1V_1C_2V_2$, where the medial consonant C_2 was one of the nasals [m] or [n], and the vowel pattern V_1–V_2 was one of [a]–[i], [a]–[i], [u]–[a] or [i]–[a]. All possible combinations of onset, medial consonant and vowel pattern were recorded by a male native English speaker.

Participants were divided into three groups. Each group was assigned to one of the languages (Control, Single Type or Two Types). The exposure phase in all languages included six FILLER words, two starting with each of the onsets [w], [y] and [l]. Participants who were taught the Control language were only exposed to these six control words (see Table 6.4c). The Single Type language additionally included two words starting with the same CONF onset ([p], [t] or [k], counterbalanced across participants; see Table 6.4a). Finally, the exposure phase in the Two Types language included two words, each starting with a different CONF onset ([p] and [t], [p] and [k], or [t] and [k], counterbalanced across participants), in addition to FILLER words (see Table 6.4b). All participants received a single exposure set.

Exposure words with FILLER onsets were included to make the exposure phase longer; without them the exposure phase of the Single Type language would have been only two words long. Approximants such as [w], [y] and [l] are considered to be voiced, though they are neither stops nor fricatives (Hayes, 2011). If anything, they should provide support for the voiced fricative test onsets (NONCONF-UNATT) rather than the voiceless stop ones (CONF-ATT); any preference for CONF-UNATT over NONCONF-UNATT test onsets, then, would be despite rather than because of the FILLER onsets.

In the test phase, all participants rated five novel words, one with each of the five onsets [p], [t], [k], [z] and [ð]. For consistency, we refer to [p], [t] and [k] as CONF test onsets and to [z]

and [ð] as NONCONF test onsets in all three languages, even though one of them, the Control language, didn't provide any evidence for the generalization. Since none of the languages had NONCONF onsets in the exposure phase, NONCONF onsets were always unattested (NONCONF-UNATT). The exposure phase of the Control language didn't have any CONF onsets; [p], [t] and [k] were therefore all CONF-UNATT. The test phase of the Single Type language had one CONF-ATT and two CONF-UNATT onsets, and the test phase of the Two Types language had two CONF-ATT and one CONF-UNATT onsets.

6.5.1.2 Participants

A total of 450 participants were recruited through Amazon Mechanical Turk: 50 participants in each of the three lists for the Single Type and Two Types languages, and 150 participants in the Control language. Nine participants were rejected because they reported that English was not their only native language. We report data from the remaining 441 participants (233 women, 204 men, four unreported; median age: 28, age range: 18-71, one unreported).

6.5.1.3 Statistical analysis

The statistical analysis was similar to previous experiments, with the exception that our design did not allow us include an onset type random slope for participants, since we only had a single observation per participant for some of the combinations of onset category and language (e.g., there was only one test token with a CONF-UNATT onset in the Two Types language). As such, the random effect structure in all LMEMs reported below only included random intercepts for subjects and for onsets.

6.5.2 Results

Figure 6.4 shows the mean endorsement rates for each onset type in each of the languages. The design was not fully crossed due to the absence of CONF-ATT onsets from the test phase of the Control language. Consequently, we performed two separate analyses: one that included all three

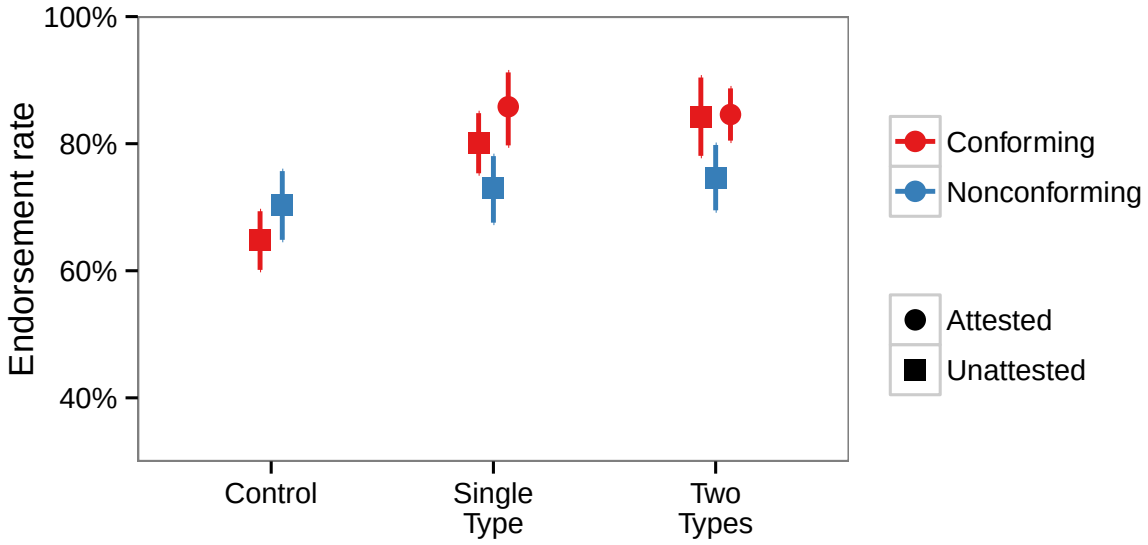


Figure 6.4: Mean endorsement rates for Experiment 3. Error bars represent bootstrapped 95% confidence intervals.

language, but only test words with CONF-UNATT and NONCONF-UNATT onsets; and another that included all three onset types, but only the Single Type and Two Types languages.

6.5.2.1 Excluding test words with CONF-ATT onsets

The two main effects were not significant (language: $\chi^2(2) = 1.74$, $p = .42$; onset type: $\chi^2(1) = 1.16$, $p = .28$), but the interaction between language and onset type was significant ($\chi^2(2) = 13.32$, $p = .001$). This interaction was driven by higher endorsement rates for CONF-UNATT than NONCONF-UNATT onsets in both the Single Type and Two Types languages (Single Type: $\chi^2(1) = 4.11$, $p = .04$; Two Types: $\chi^2(1) = 5.99$, $p = .01$), but not in the Control language, where there was a nonsignificant trend in the opposite direction ($\chi^2(1) = 2.68$, $p = .1$).

The significant simple effect in the Single Type language confirms that learners generalized based on a single CONF onset type in exposure. The reverse trend in the Control language may reflect a tendency to interpret the approximant FILLER onsets in exposure as providing support for voiced over voiceless onsets.

6.5.2.2 Excluding the Control language

The main effect of onset category was significant ($\chi^2(2) = 12.58$, $p = .002$); the main effect of language was not significant ($\chi^2(1) = 0.13$, $p = .72$), and neither was the interaction ($\chi^2(2) = 1.22$, $p = .54$). This indicates that the pattern of results is not statistically different across the Single Type and Two Types language.

To further examine the effect of onset category, we performed pairwise comparisons across the levels of this factor. The difference in endorsement rate between CONF-UNATT and NONCONF-UNATT was significant ($\chi^2(1) = 8.03$, $p = 0.005$) and did not interact with language ($\chi^2(1) = .39$, $p = .53$). There was no significant difference between test words with CONF-ATT and CONF-UNATT onsets ($\chi^2(1) = 1.55$, $p = .21$), and again no interaction with language ($\chi^2(1) = 1.22$, $p = .27$).

Finally, we assessed the statistical significance of the difference between words with CONF-ATT and CONF-UNATT onsets within each language separately. Endorsement rates within the Two Types language were indistinguishable ($\chi^2(1) = 0$, $p = .96$); conversely, there was a trend towards higher ratings for CONF-ATT than for CONF-UNATT onsets within the Single Type language ($\chi^2(1) = 2.79$, $p = .09$).

6.5.3 Discussion

Specific-to-general models predict that generalizations cannot be formed until at least two specific instantiations of the generalization have been encountered. Experiment 3 tested this prediction by exposing participants to the Single Type language, in which two tokens of a single type of voiceless stop onset—e.g., [p]—supported the generalization that onsets can be voiceless stops. Contrary to the prediction of specific-to-general models, participants generalized to unattested generalization-conforming onsets (CONF-UNATT; e.g., [t]), preferring them over NONCONF-UNATT onsets such as [z].

The Control language was designed to rule out two interpretations of the preference that participants who learned the Single Type language showed for CONF-UNATT over NONCONF-UNATT

onsets: first, that participants had a prior preference for voiceless stops, either due to statistical patterns in the English lexicon or for any other reason; and second, that the preference was due to the presence of six approximant FILLER onsets (though this scenario is unlikely given the absence of any shared phonological features between approximants and voiceless stops). After exposure to this language, which included only FILLER onsets, participants rated test words with CONF-UNATT onsets *lower* than ones NONCONF-UNATT onsets, suggesting that the pattern of endorsement rates for the Single Type language were indeed due to generalization from the single type of voiceless stop in the exposure phase.

In a third language, the Two Types language, the generalization was supported by one token each of two different types of voiceless stops, e.g., both [p] and [k]. Participants again generalized to test words with a CONF-UNATT onset; moreover, they did not distinguish CONF-UNATT from CONF-ATT onsets, replicating the One Set group of Experiments 1 and 2a. There was no clear evidence of a preference for CONF-ATT over CONF-UNATT in the Single Type language either, though there was a suggestive numerical difference. The two languages differed in that the Two Types language had a single token of each of the two types of voiceless stop, whereas the Single Type language had two tokens of the same type. While this decision served to equalize the number of exposure words across the languages, two tokens of the same onset appear to be sufficient for some onset-specific learning (compare the Two Sets group of Experiments 1 and 2a), which may explain the numerical trend in the Single Type language.

6.6 General discussion

Speakers generalize their phonotactic knowledge from the sounds attested in their language to sounds from the same phonological classes as the attested ones. If words often end with voiceless stops [p] and [k], speakers will judge words that end with other voiceless stops (e.g., [t]) to be acceptable. The experiments presented in this paper investigated the process by which such generalizations are acquired, focusing on evaluating whether learning follows a specific-to-general

order: Do multiple specific instantiations of a generalization need to be acquired before the generalization can be formed?

In Experiments 1 and 2a, participants were divided into groups that received varying amounts of exposure to an artificial language. The consonants of the words in each of the languages supported a phonotactic generalization, but not all possible consonant patterns that conformed to the generalization were presented to participants. In both experiments, participants learned the abstract generalization following minimal exposure: When asked to rate novel words for acceptability, participants in all groups judged words that conformed to the generalization to be better formed than words that did not. By contrast, participants did not start distinguishing the specific sounds they were exposed to from the ones they were not exposed to until they received additional exposure to the language. In other words, participants showed evidence of learning a generalization (e.g., that consonants are often identical) before they showed evidence of learning any of its specific instances (e.g., that [p, p] is a valid consonant pair).

In both Experiments 1 and 2a, the generalization that participants learned was supported by multiple types. In the critical condition of Experiment 3, by contrast, participants were only exposed to a single type of consonant which represented a phonological class. Even when the amount of exposure to the generalization was thereby reduced to the absolute minimum, participants still generalized to other members of that class.

The finding that participants can generalize from a single type is most naturally consistent with simultaneous learning models, in which learners can form generalizations about classes of sounds at the same time as (or before) they acquire knowledge about specific sounds from those classes. Maximum entropy models are a prominent implementation of the simultaneous learning view (Hayes & Wilson, 2008; Pater & Moreton, 2012). In these models, the well-formedness of a sound is derived from a linear combination of weights associated with each of the phonological classes that contain the sound. The well-formedness of a [b] is determined by both the weight for [b] and the weight for the class of voiced stops (as well as the various other classes that [b] belongs to). Learning the phonotactics of the language consists in determining the set of weights that is

most consistent with the statistical distribution of sounds in the language. Since both the sound-specific and the class-wide weight contribute to the well-formedness of a [b] token, exposure to this token will cause both weights to be increased. Even if the only token the learner was exposed to was a [b], then, the learner will rate a novel voiced stop such as [g] as better than a voiceless one such as [k].

After a single exposure set in Experiments 1 and 2a, as well as in the Single Type and Two Types languages of Experiment 3, there was no statistically significant difference between sounds that conformed to the broad generalization and were attested in exposure and those that conformed to the generalization but were not attested in the exposure phase; by contrast, the difference between conforming to nonconforming sounds as a whole was consistently significant. While maximum entropy models predict rapid generalization to unattested conforming sounds, they do not predict that all conforming sounds should be judged as equally well formed. It is true that a single exposure to a [b] would lead a maximum entropy learner to increase some of the weights that apply to [d] (e.g., the weight for voiced stops); at the same time, weights that apply to [b] but not to [d], such as the weight for labials or a weight specific to [b], would also be increased. Consequently, conforming attested sounds (here [b]) would be preferred to conforming unattested ones ([d]). The prediction of both a generalization *and* and an attestation effect is consistent with the empirical endorsement rates after multiple exposure sets, but is inconsistent with the pattern that emerged after minimal exposure (with the usual caveats about the interpretation of negative results).

The absence of an attestation effect after limited exposure can be characterized as a *general-to-specific* learning pattern: generalizations over classes of sounds emerge before generalizations about specific sounds. As Linzen and O'Donnell (2015) show, the general-to-specific empirical pattern does not necessarily reflect an explicit general-to-specific learning strategy. Instead, it can emerge from a simultaneous learner that incorporates a parsimony bias encouraging the learner to represent the input using a single generalization (cf. Chomsky & Halle, 1968, p. 337). For example, after exposure to five different types of voiced onsets (as in Experiment 1), this bias may lead participants to characterize words in the language as beginning with voiced consonants—a single

generalization—rather than as beginning with [g], [b], [v], [z] or [ð] (five separate generalizations). As learners receive more exposure to the language, however, the absence of conforming unattested sounds becomes more apparent, and prompts learners to revert to a less parsimonious but more accurate sound-specific representation.

The rapid generalization results of Experiment 1 and 2a can be reconciled with the specific-to-general view if we assume that learners always record the identity of the specific sounds they have been exposed to, but avoid deploying that knowledge if the number of exposure words that contained that sound was below a certain threshold. Under this assumption, knowledge about specific sounds might give rise to generalizations about classes of sounds, but would not necessarily lead to a difference in acceptability between attested and unattested sounds. While existing specific-to-general models of phonotactics do not incorporate this assumption—in both the Minimal Generalization Learner (Albright, 2009) and StaGe (Adriaans & Kager, 2010), a generalization cannot have a stronger effect than all of its specific instances—these models can be modified to include a component that discounts knowledge based on fewer tokens (Albright & Hayes, 2002; for simulations, see Linzen & Gallagher, 2014).

It is harder to see how the specific-to-general view could be reconciled with the single-type generalization pattern observed in Experiment 3. Indeed, it seems implausible that a learner that has been exposed to one or two words that start with a [b] will conclude that all words in the language start with a [b] until shown evidence to the contrary. One could imagine a model where upon hearing a word that starts with a particular sound the learner instantaneously constructed all generalizations that contain that sound, of course, but that learner would be empirically indistinguishable from a simultaneous one.

The specific-to-general assumption in phonotactic learning is an example of a broader conservative generalization strategy that is often attributed to language learners. The conservative generalization assumption is motivated by the so-called "subset problem". The form of the argument is as follows (Dell, 1981). The goal of the learner is to characterize the grammar that generated the input it is observing. It can only use positive evidence (i.e., attested forms): No one

tells a learner of English that **mpepm* is phonotactically illegal. Suppose that the onsets that the learner has been exposed to are [b], [d] and [g]. This input is compatible with the following two grammars (among others): in Grammar 1, all words start with a voiced stop; in Grammar 2, all words start with a stop (either voiced or voiceless). The language generated by Grammar 1 is a subset of the language generated by Grammar 2. If at one point in the learning process the learner selects Grammar 1, and later on encounters a word that starts with a voiceless stop (e.g., [k]), the learner can revise its decision and assume the wider Grammar 2 instead. The reverse decision is argued to be impossible because of the absence of negative evidence: A learner that chose Grammar 2 would never receive evidence that the generalization was too wide.[5] To avoid overly broad generalizations, learners have to be conservative: "Whenever there are two competing grammars generating languages of which one is a proper subset of the other, the learning strategy of the child is to select the less inclusive one" (Dell, 1981, p. 34). This strategy was later termed the Subset Principle (Berwick, 1985; M. Hale & Reiss, 2003).

The subset problem only arises in a specific theoretical model of language acquisition, however (Gold, 1967; Berwick, 1985), and its practical relevance has been called into question (A. Clark & Lappin, 2011). In particular, while it is true that learners rarely (if ever) receive direct evidence that certain sound combinations are impossible in their language, they often receive *indirect* negative evidence in the form of frequency asymmetries. Suppose that the learner is exposed to a language in which words start with either [b] or [d] (a simplified version of Experiment 1). After encountering two words in the language, one that started with [b] and one that started with [d], the learner might conclude that the best characterization of the phonotactics of the language is that all words start with voiced stops. As the learner receives additional exposure to language, however, the systematic absence of [g] onsets becomes more and more conspicuous. If words could start with any

[5] In fact, under the assumption that simpler grammars—grammars that can described more succinctly—are preferred to complicated ones (Chomsky & Halle, 1968), a learner would typically select the *widest* grammar possible, unless it is equipped with a countervailing bias such as the Subset Principle (Dell, 1981).

voiced stop, the absence of this particular voiced stop would become a "suspicious coincidence" (Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). This constitutes indirect negative evidence that may cause the learner to revert to a narrower hypothesis, obviating the need for the Subset Principle. This strategy is implemented in the probabilistic model proposed by Linzen and O'Donnell (2015).

Our findings are likely to be applicable outside the domain of phonotactic learning proper. The pronunciation of morphemes often depends on the phonological environment surrounding them. The regular English past tense morpheme *-ed*, for example, has three different variants (allomorphs), [ɪd], [t] and [d]. The choice of allomorph in each case is determined by the final sound of the verb: [ɪd] is used after [t] and [d] (*tended* [tɛndɪd]), [t] after voiceless consonants (*wrapped* [ɹæpt]), and [d] elsewhere (*banned* [bænd]). Speakers use these phonological generalizations to form the past tense of novel verbs: even a speaker who is unfamiliar with the verb *schlep* will agree that its past tense is formed with the [t] rather than the [d] allomorph. This also applies to novel sounds: an English speaker who pronounced the last consonant of *Bach* using the original German [x], which is a voiceless consonant, would likely form the past tense of a putative verb *out-Bach* as [bɑxt] rather than [bɑxd] or [bɑxɪd] (Pinker, 1999).

It is natural to assume that generalizations related to allomorph selection are formed in the same way as the phonotactic generalizations discussed in the current paper (Albright & Hayes, 2003; Gouskova, Newlin-Łukowicz, & Kasyanenko, 2015). The predictions made by specific-to-general models in the case of limited exposure to the language are perhaps even less plausible in the morphological context than in phonotactics. Consider, for example, a learner of English who has only been exposed to the past form of a handful of words that happen to all end with [k], and therefore form their past form with a [t] (e.g., *kicked* and *talked*). This learner would be completely stumped as to the past form of *stop*: it wouldn't even be able to make a guess, since no generalization would have been formed (for a similar point, see Kapatsinski, 2014, p. 16). The results of the present study lead us to predict that learners will show simultaneous rather than specific-to-general learning in the allomorphy scenario as well.

This paper has advocated for a learning procedure that includes phonological classes at all levels of generality as part of the learner's hypothesis space. The existence of phonological classes in the learner's hypothesis space raises the question of the origin of those classes: How do learners know that [b], [d] and [g] form the class of voiced stops, but [b], [s] and [m] do not constitute a phonological class to the exclusion of other sounds? Various answers have been proposed to this question. The inventory of phonological features may be innate (Chomsky & Halle, 1968; M. Hale & Reiss, 2003); phonological features may emerge as learners group together sounds that sound similar to each other or are articulated in a similar way (Lin & Mielke, 2008); finally, they may simply reflect groupings of sounds according to their behavior in phonological alternations (Mielke, 2008). Our results do not bear on this debate, for two reasons. First, this debate concerns the origin of phonological features in infants. Participants in our experiments were all English speakers, and all of the sounds in the artificial languages they learned were drawn from the English inventory. Consequently, their hypothesis was likely based on the features that are in active use in English phonology.

Second, the question of the origin of phonological features is orthogonal to the distinction between specific-to-general and simultaneous learning more generally. In order to form a generalization over two specific sounds, say [b] and [g], the learner must already be able to represent the fact that the two sounds belong to the same phonological class (specifically, voiced stops); this is exactly the same knowledge that a simultaneous learner would require to analyze a single instance of [b] as a voiced stop (M. Hale & Reiss, 2003). All extant models of generalization in phonotactics, then, assume that the learner is equipped with the ability to represent phonological classes. It is conceivable, of course, that infants would exhibit two distinct developmental stages, one in which statistics about specific segments were accumulated, and another in which generalizations to wider phonological classes were formed. The empirical evidence to date points in the opposite direction, however (Saffran & Thiessen, 2003; Cristia & Peperkamp, 2012). In reality, the acquisition of phonological features, phonotactics and lexical knowledge may well be intertwined (Feldman, Griffiths, Goldwater, & Morgan, 2013).

All of our participants were English-speaking adults. As such, our experiments are a closer approximation of second language learning than of first language acquisition. At the same time, we are encouraged by the fact that our findings are convergent with results from infant studies. Six month old infants exposed to a language very similar to the one used in Experiment 1 showed a similar behavior to the adult participants in the One Set group of Experiments 1 and 2a: They looked longer at words that started with CONF-UNATT than NONCONF-UNATT onsets, but did not distinguish CONF-UNATT from CONF-ATT onsets (Cristia & Peperkamp, 2012). In another experiment, nine-month-olds who have been exposed to a single word with a duplicated syllable (*leledi*), repeated a few times, preferred novel words with a similar structure, suggesting that they learned a reduplication rule from a single example (Gerken, Dawson, Chatila, & Tenenbaum, 2015); this is consistent with the finding of single-type generalization in Experiment 3.

6.7 Conclusion

This paper reported on a series of artificial language experiments of phonotactic generalization in artificial language learning. The experiments showed that participants can learn both segment-specific and abstract phonotactic patterns in an artificial language following a very short exposure session. Abstract patterns (e.g., word onsets are voiced) were learned more quickly than segment-specific ones (e.g., [b] and [d] are valid word onsets); this applied regardless of whether the abstract pattern was categorical or probabilistic, and of whether it was based on a phonological class or on an identity relation across segments. Finally, abstract patterns were acquired on the basis of a single example. We conclude that humans are not conservative generalizers; consequently, models of phonotactic learning should not be based on the assumption that knowledge about specific items is a prerequisite for the formation of abstract generalizations.

7 A model of rapid phonotactic generalization

7.1 Introduction[1]

Chapter 6 presented empirical evidence that humans show rapid phonotactic generalization. This chapter describes a computational model of the findings: a probabilistic phonotactic learner that acquires patterns at multiple levels of generality. In particular, it can acquire a broad pattern based on a single example. The model trades off the fit to the learning data with two simplicity pressures: first, a preference for phonological classes that are specified with fewer phonological features; and second, a preference for explaining the data using as few phonological classes as possible. These pressures bias the learner towards acquiring a single broad pattern and against acquiring multiple specific patterns; however, given enough data showing that the broad pattern is inadequate, the learner will fall back to learning the specific patterns.

7.2 The model

We first focus on the generation of a single sound (e.g., simple syllable onsets): sections 7.2.1 and 7.2.2 describe the generative process in the case that all of the sounds in the input are assumed to belong to the same phonological class; we then relax this assumption and allow the sounds to come

[1] Parts of this model were developed in collaboration with Timothy O'Donnell.

from one of several possible phonological classes (section 7.2.3). Section 7.2.4 describes how the model is extended to generate pairs of sounds.

The model describes a probabilistic process that leads to the generation of a sound or a sequence of sounds. The goal of the learner is to invert that process: given the sounds that the learner has heard, what are the parameters of the generative process likely to be? For example, how many phonological classes are there, and what are they? Section 7.2.6 describes the inference process.

7.2.1 Prior over phonological classes

We first describe the model's prior probability distribution over phonological classes. In the notation we use, features are sets that contain the possible values of the feature, e.g.:

$$\textit{voice} = \{\text{+}\textit{voice}, \text{-}\textit{voice}\} \tag{7.1}$$

$$\textit{place} = \{\textit{labial}, \textit{coronal}, \textit{dorsal}\} \tag{7.2}$$

F is the set of all such features, i.e. $F = \{\textit{voice}, \textit{place}, \ldots\}$. We define a specification probability p, which controls the model's willingness to consider highly specified classes: low values of p strongly encourage underspecified classes, such as [] or [+voice], whereas high values of p favor highly specified classes, such as [+voice, labial, -continuant]. This probability will later be estimated from the learning data. Given a particular value of p, the prior probability over phonological classes is obtained from the following simple probabilistic process that generates a class c:

- For each feature $f_i \in F$:
  - With probability p, include f_i in the specification of the class and select a value $v_i \in f_i$ for the feature uniformly.
  - Otherwise, leave f_i unspecified.

Suppose, for example, that $p = 0.4$ and the language only has three phonological features, $F = \{\textit{voice}, \textit{place}, \textit{continuant}\}$. Some examples of the prior probability distribution over classes are:

- The prior probability of the empty class [], which includes all segments, is the probability of leaving all three features unspecified:

$$P([]) = (1-p)^3 = 0.6^3 = 0.216 \tag{7.3}$$

- The class [+voice] has one specified feature, *voice*, and two unspecified ones, *place* and *continuant*. The probability of this pattern of specification is

$$p \times (1-p)^2 = 0.6^2 \times 0.4 = 0.144 \tag{7.4}$$

  There are two possible classes for which only *voice* is specified: [+voice] and [-voice]. The probability of each is $0.144/2 = 0.072$.

- The class $c_0 =$ [+continuant, labial] will be generated if *voice* is left unspecified (this happens with probability $1-p$) but *continuant* and *place* are not (probability p each). Finally, [*labial*] needs to be selected out of the three values for *place* and [+*continuant*] out of the two values for *continuant*. The probability assigned to c_0 is therefore

$$P(c_0) = p^2(1-p) \times \frac{1}{2} \times \frac{1}{3} = 0.016 \tag{7.5}$$

While for most values of p less specified classes are indeed less likely, for p values very close to 1 the reverse might be the case; for example, if $p = 0.9$, we get

$$\begin{aligned} P([]) &= 0.1^3 = 0.001 \\ P([+voice]) &= 0.9 \times 0.1^2 \times 0.5 = 0.0045 \\ P([+continuant, labial]) &= 0.9^2 \times 0.1 \times \tfrac{1}{2} \times \tfrac{1}{3} = 0.01 \end{aligned} \tag{7.6}$$

In the two artificial languages that we consider, the empirical estimates of p are much lower than that, so it is unclear whether this ever becomes a concern in practice.

7.2.2 Segment generation

We assume that each of the segments in the class has the same probability (cf. the "size principle" of Tenenbaum & Griffiths, 2001). Consider again the class [+continuant, labial]. Under the assumption that the model's segment inventory is the English inventory (see Table 7.1 below), there are only two segments that are labial continuants, [v] and [f]; the probability of each one of them being generated from c_0 will be $P(s|c_0) = 1/2$.

7.2.3 Mixture of components

In a realistic scenario, the sounds of the language are likely to be generated from more than one phonological class. For example, voiced stops and voiceless fricatives may both be acceptable onsets. The classes may well be at different levels of generality: the learner may want to encode knowledge about voiced stops in general and about one particular voiced stop (e.g., [b]). In this section we extend the model to be a mixture of distributions of the type described in the previous section.

Since we cannot assume that the learner knows in advance how many mixture components there are, we use the nonparametric Dirichlet process mixture model (see e.g. Gershman & Blei, 2012 for a review). A Dirichlet process can be represented as a prior over partitions of the data (the Chinese restaurant process). This prior favors grouping the segments into fewer components (natural classes); among partitions with the same number of components, it tends to favor skewed

partitions where a handful of the components accounts for much of the data (the "rich get richer" property). The first property constitutes an intuitively desirable parsimony bias. It is less clear whether there is a straightforward motivation to the second property, though many linguistic phenomena do have the property that a handful of underlying types are responsible for a majority of the tokens, e.g. word frequency distributions (Goldwater, Johnson, & Griffiths, 2006). The Dirichlet process prior has a parameter α called the concentration parameter; the lower α is, the stronger the bias to have fewer components.

As an example, we consider the case of five data points (segments) and compare the prior probability of different partitions (with α set to 1). We use the notation $P(\langle 3, 1, 1 \rangle)$ to indicate the probability of a partition that has three components, one of which with three segments and the other two with one segment each.[2]

- $P(\langle 5 \rangle) = 0.2$
- $P(\langle 4, 1 \rangle) = 0.05$
- $P(\langle 3, 2 \rangle) = 0.0166$
- $P(\langle 1, 1, 1, 1, 1 \rangle) = 0.008$

The prior probability of a partition where all of the tokens are generated from the same natural class is thus more than 20 as high as the prior probability of each one of them coming from a different natural class; this illustrates the parsimony bias.

[2]To calculate these probabilities, I used the following formula (N is the number of segment tokens, K is the number of components, n_k is the number of segment tokens that are assumed to have been generated by the k-th component; see Gershman & Blei, 2012):

$$\frac{\alpha^k \prod_{k=1}^{K} (n_k - 1)!}{\prod_{i=1}^{N} (i - 1 + \alpha)} \tag{7.7}$$

This prior can be constructed as a process. This representation will be useful when we will need to derive predictions for new segments from the model (a predictive distribution). Suppose that s_i is an ordering of the input sounds, and that we know which components generated the first $n-1$ sounds $s_1, \ldots, s_{n-1}$. If K is the number of mixture components that have been posited so far and $n_1, \ldots, n_K$ is the number of sounds that have been drawn from each mixture component, then the probability distribution that specifies which component z_n will be the component that s_n will be drawn from is given by

$$P(z_n = k | z_{1:n-1}) = \begin{cases} \dfrac{n_k}{n-1+\alpha} & \text{if } k \leq K \\ \dfrac{\alpha}{n-1+\alpha} & \text{otherwise} \end{cases} \tag{7.8}$$

where higher values of the concentration parameter α encourage the learner to posit more components, and lower values lead to a more conservative behavior. In other words, the learner assumes that the next segment it will get belongs to one of the existing classes with a probability that approaches 1 as more and more tokens are observed, though it always reserves a small amount of probability mass to the option of adding a new phonological class to the current inventory. Given that the next segment indeed comes from one of the existing class, the probability that a particular class is the generator is proportional to the number of segments currently assigned to that class (hence the "rich get richer" property).

7.2.4 Sequences and spreading

So far our component parameters have been individual natural classes. To simulate Experiment 2, the parameters will need to be *pairs* of natural classes. Moreover, we need a special representation for a sequence of two identical consonants: sequences of two identical voiced stops include ([b], [b]) and ([g], [g]), but not ([g], [b]). To accomplish this, we assume that a sequence can have either two independent classes, e.g. ([+voiced], [labial, -cont]), or a single class and a vacant slot, e.g. ([+voiced], X); whenever a pair specification contains a vacant slot, the segment in the previous

position automatically spreads into the slot (Colavin, Levy, & Rose, 2010). The procedure for generating a class sequence is as follows:

1. Draw a class c_1 from the distribution defined in section 7.2.1.
2. Draw $Q \sim \text{Bernoulli}(q)$.
3. If $Q = 1$, return the sequence (c_1, X).
4. Otherwise, draw an additional class c_2, and return (c_1, c_2).

To generate segments from a pair of classes we again sample uniformly from all pairs of segments included in each pair of classes. The class pair (c_1, c_2) includes all segment pairs (s_1, s_2) such that $s_1 \in c_1$ and $s_2 \in c_2$. For example, ([+voiced], [-voiced]) includes [bk] and [vt], among other segment pairs.

The class pair (c, X) includes all segment pairs (s, s) such that $s \in c$ (and only those pairs). In other words, the same segment s needs to be repeated twice, rather than each segment being an independent sample from c. So while ([+cont, labial], [+cont, labial]) includes [ff], [fv], [vf] and [vv], ([+cont, labial], X) only includes [ff] and [vv].

7.2.5 Hyperparameters

We place the following priors on the hyperparameters:

$$\begin{aligned} p &\sim \text{Beta}(1, 1) \\ \alpha &\sim \text{Gamma}(2, 4) \end{aligned} \tag{7.9}$$

When we simulate Experiment 2, we place a flat beta prior on q:

$$q \sim \text{Beta}(1, 1) \tag{7.10}$$

7.2.6 Inference

The goal of the learner is to infer the parameters of the generative model given a list of sounds. Inference follows Bayes' law:

$$P(\theta|D) \propto P(\theta)P(D|\theta) \tag{7.11}$$

where D is the learning data and θ is a vector of all parameters – the hyperparameters, the partition of the segments into groups and the natural classes assumed to generate each group of segment. $P(\theta)$ here is the prior probability of the parameter vector θ, and $P(D|\theta)$ is the likelihood of θ given the data. Our objective here isn't to find the single set of parameters that has the highest posterior probability, but rather to infer the full posterior probability distribution over hypotheses.

We illustrate the goals of the inference process on the training set $D = \{\text{b, d, g, v, ð}\}$. Consider the hypothesis θ_1, in which $p = 0.5$, $\alpha = 1$, all of the segments are assumed to come from the same natural class, and the identity of that natural class is [+voice]. The prior probability of this partition hypothesis is

$$P(\theta_1) = P_{Beta(1,1)}(0.5)P_{Gamma(2,4)}(1)P(\langle 5 \rangle)P([+voice]) \tag{7.12}$$

where $P_{Beta(1,1)}$ and $P_{Gamma(2,4)}$ are the probability density functions of the respective distributions, $P(\langle 5 \rangle)$ is the prior probability of the partition that groups all segments into one class, and $P([+voice])$ is the prior probability of the natural class [+voice]. If the learner assumes the English segment inventory (Table 7.1), there are 11 segments with the feature [+voice]; the likelihood contributed by each segment is then $\frac{1}{11}$, and the total likelihood is

$$P(D|\theta_1) = \left(\frac{1}{11}\right)^5 \tag{7.13}$$

Compare this to the hypothesis θ_2, in which the hyperparameters are the same ($p = 0.5$, $\alpha = 1$), but the segments are assumed to each come from a different natural class that picks them out

uniquely, e.g. [+voice, -continuant, labial] for [b]. The prior probability of this hypothesis is lower, for several reasons: first, since $P(\langle 1,1,1,1,1\rangle) < P(\langle 5\rangle)$ (see section 7.2.3); second, since multiplying the probabilities of several natural classes will in general result in a lower probability than having just one natural class; and third, because each of the natural classes are more richly specified than in θ_1. At the same time, the likelihood of θ_2 is higher; once one of the five cluster has been selected–which happens uniformly (see Equation 7.8)–the identity of the segment is deterministic, because the classes each contain exactly one segment. The total likelihood is therefore

$$P(D|\theta_1) = \left(\frac{1}{5}\right)^5 \tag{7.14}$$

If each of the segments is repeated n times in the training set, the difference in likelihood between the two hypotheses will increase to $(1/11)^{5n}$ vs. $(1/5)^{5n}$. Since the prior probability of the hypotheses remains the same (by definition), for a high enough n any advantage that θ_1 has over θ_2 in terms of prior probability will be offset by a disadvantage in likelihood.

Even setting aside the issue of the continuous prior distributions over hyperparameters, the space of all possible combinations of partitions and natural classes is enormous, even for a small training set; this makes it unfeasible to enumerate all such combinations. We therefore use Markov chain Monte Carlo (MCMC) to sample from the posterior probability. Specifically, we use a Gibbs sampler for the Dirichlet process mixture model (Neal, 2000). After each Gibbs sweep, we sample the concentration parameter α using the auxiliary variable method of Escobar and West (1995); we then use slice sampling (Neal, 2003) to sample the specification probability p and spreading probability q (when relevant).

We run the sampler for 3000 iterations. We discard the first 100 samples, and keep every fifth sample of the remaining samples. This gives us 580 samples from the posterior distribution over hyperparameters, partitions of the data into clusters and cluster parameters. Recall that by contrast with many other models our goal is not to find the mode of the posterior distribution, i.e. a single hypothesis that fits the data best. Instead, we assume that the learner maintains the full posterior probability distribution over hypotheses and uses the full posterior to make predictions

segment	continuant	sonorant	voice	place	distributed	anterior	nasal
p	-	-	-	labial	0	0	-
b	-	-	+	labial	0	0	-
m	-	+	+	labial	0	0	+
f	+	-	-	labial	0	0	-
v	+	-	+	labial	0	0	-
θ	+	-	-	coronal	+	+	-
ð	+	-	+	coronal	+	+	-
t	-	-	-	coronal	-	+	-
d	-	-	+	coronal	-	+	-
s	+	-	-	coronal	-	+	-
z	+	-	+	coronal	-	+	-
n	0	+	+	coronal	-	+	+
l	0	+	+	coronal	-	+	-
ʃ	+	-	-	coronal	+	-	-
ʒ	+	-	+	coronal	+	-	-
r	0	+	+	coronal	+	-	-
k	-	-	-	dorsal	0	0	-
g	-	-	+	dorsal	0	0	-

Table 7.1: Features used in the phonotactic mixture model simulation.

about the probabilities of a new segment. Given an individual sample from the posterior, which includes both a clustering $z_{1:n}$ of the sounds into components and the phonological class $c_{1:K}$ that parametrizes each component, the predictive distribution (the distribution of the next sound that will be generated from the mixture model) is given by

$$P(s_{n+1} = s_0) = \frac{\alpha}{n+\alpha} P(s_0) + \sum_{k=1}^{K} \frac{n_k}{n+\alpha} P(s_0|c_i) \tag{7.15}$$

To approximate the full posterior predictive probability of a segment, then, we calculate (7.15) for each of the samples from the posterior, and average those probabilities.

7.3 Simulation: Experiment 1

7.3.1 Procedure

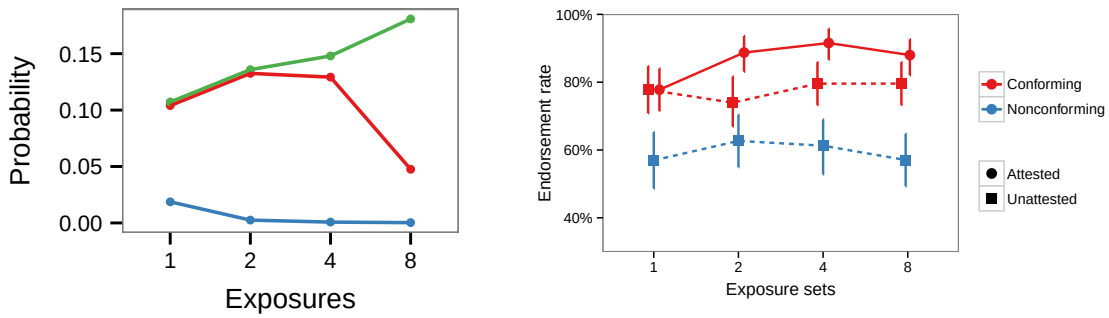
The input to the model consisted of the onsets of the training words: the model was not required to parse the words into syllables. The model was trained on the same 12 languages that the human participants were exposed to, in the same four exposure conditions (1, 2, 4 and 8 exposures). The results presented below are averaged across all 12 languages.

The feature inventory was based on Hayes (2011), with the following simplifications aimed to reduce the number of features: first, we only included features that are distinctive in English (this modification is also consistent with the fact that our goal is to simulate a native English speaker); second, *place* was represented as a single ternary feature, whose values can be [labial], [dorsal] and [coronal], rather than as three separate binary features; and third, only consonant features were included. The full inventory is shown in Table 7.1.

7.3.2 Results

Figure 7.1 shows the posterior predictive distribution: the probabilities that the next segment generated from the model will be one of the attested, legal or illegal segments, respectively. After fewer exposures, the learner does not distinguish attested from unattested segments within the legal class. As the model receives more input it homes in on the particular sounds it encountered in training and becomes less likely to generalize.

We now inspect the posterior distribution over phonological classes. Following Equation 7.15, we define the posterior probability of a class c as the average proportion of input sounds that were assigned to a component that had c as its parameter (note that multiple components can have the same parameter). Table 7.2 shows the classes that had the highest posterior probability after 1, 2, 4 and 8 exposures, for one of the 12 lists. The same set of sounds is often specified in several different ways; for example, voiced stops can be specified either as [-sonorant, +voice] or as [-sonorant, +voice, -nasal]. After one exposure, the classes are fairly broad; they become narrower as the



(a) (b)

Figure 7.1: (a) Simulation of Experiment 1: Predictive probability in the three test conditions, for words with onsets encountered in training (Attested), onsets not encountered in training but similar in voicing to training onsets (Legal) and onsets that differed in voicing from training onsets (Illegal), averaged across the 12 experimental lists. (b) Endorsement rates from Experiment 1.

learner receives more exposures. Note that even after eight exposures, the model does not always prefer classes that include exactly one segment: whenever a class exists that includes multiple attested segments and no unattested segments, those classes will typically be preferred over more specific ones, due to the penalty on specifying multiple features.

Figures 7.2a and 7.2b show the posterior distribution of the hyperparameters after one and eight exposures respectively. After one exposure, the learner typically prefers to assume that all of the segments come from a single class. This class will tend to have a small number of specified features (e.g. [+voiced] or even []). This leads the model to conclude that the specification probability in the language is relatively low. After eight exposures the learner posits more classes (though, as mentioned above, less classes than attested segments); these classes are more specific, leading the learner to conclude that the specification probability is higher.

7.3.3 Comparison with behavioral data

Figure 7.1b repeats the results of the behavioral experiment. When tested after a single exposure to each onset, participants behave in a qualitatively similar way to the model: they distinguish Legal and Attested sounds on the one hand from Illegal sounds on the other hand, but do not distinguish

Exposures	Class specification	Sounds in class	Probability
1	[]	b, **ʒ**, **d**, ð, **f**, **k**, **m**, **l**, **ʃ**, **n**, **p**, **s**, **r**, g, **t**, **θ**, v, z	0.300
	[-sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.280
	[+voice]	b, **ʒ**, **d**, ð, **m**, **l**, **n**, g, **r**, v, z	0.173
	[-nasal, -sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.072
2	[-sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.646
	[-nasal, -sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.200
	[+voice]	b, **ʒ**, **d**, ð, **m**, **l**, **n**, g, **r**, v, z	0.058
	[-nasal, +voice]	b, **ʒ**, **d**, ð, **l**, g, **r**, v, z	0.053
4	[-sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.763
	[-nasal, -sonorant, +voice]	b, **ʒ**, **d**, g, v, z, ð	0.228
	[+voice]	b, **ʒ**, **d**, ð, **m**, **l**, **n**, g, **r**, v, z	0.002
	[]	b, **ʒ**, **d**, ð, **f**, **k**, **m**, **l**, **ʃ**, **n**, **p**, **s**, **r**, g, **t**, **θ**, v, z	0.001
8	[labial, -sonorant, +voice]	b, v	0.164
	[-nasal, labial, +voice]	b, v	0.153
	[+anterior, +continuant, +voice]	z, ð	0.139
	[-nasal, labial, -sonorant, +voice]	b, v	0.082

Table 7.2: Simulation of Experiment 1: Classes with highest posterior probability (for the list in which all training segments were voiced and [d] and [ʒ] were held out). Boldfaced characters represent sounds outside of the training set.

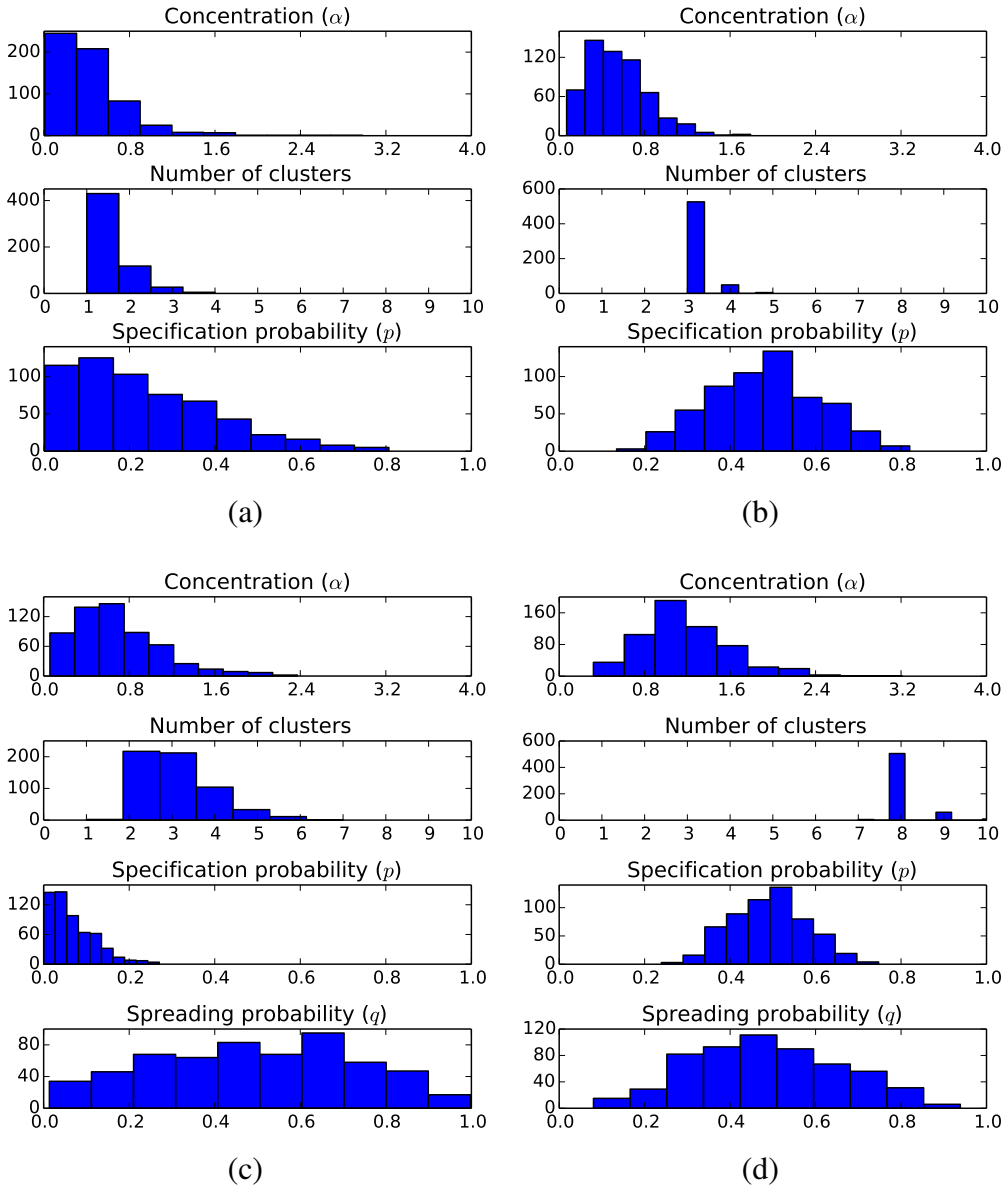


Figure 7.2: Posterior distribution over model hyperparameters: (a) Voicing experiment: 1 exposure condition; (b) Voicing experiment: 8 exposure condition; (c) Identity experiment: 1 exposure condition; (d) Identity experiment: 8 exposure condition.

Legal from Attested sounds. This suggests that they have acquired the general voicing-based phonotactic pattern, but did not acquire sound-specific patterns.

Human judgments after 2, 4 and 8 exposures showed a stable three-way distinction: Attested sounds were rated highest, Legal sounds rated lower and Illegal sounds lowest. This pattern differs from the predictions of the model in two ways: first, the model does not start showing the three-

way distinction until it has had 4 exposures to each onset, whereas humans show this distinction earlier; and second, the model converges on the particular set of attested onsets by 8 exposures, whereas humans continue to generalize to Legal onsets even after 8 exposures.

Humans rated the nonwords with Illegal onsets fairly highly (though not as highly as other types of onsets). As mentioned above, this is likely due to the fact that “illegal" test words resembled training words in all respects other than the onset: they had the same prosodic form (CVCV), the same vowels and the same middle consonants as the training words. In other words, participants trained on words with voiced onsets and tested on *timu* were comparing its acceptability to potential test words such as *aptmarʒwu*, which would presumably be rated much lower. The model does not have access to any of this information, and therefore assigns very low probability to Illegal words.

7.4 Simulation: Experiment 2

7.4.1 Procedure

The model was exposed to the consonants pairs that participants in Experiment 2a were exposed to in training; the simulation setup did not require to learn to ignore the vowels, in the same way that the simulation of Experiment 1 did not require the model to extract syllable onsets. The same phonological features were used as in the simulation of Experiment 1.

7.4.2 Results

Figure 7.3 compares the predictive distribution of the model (7.3a) to the results of the behavioral experiment (7.3b). Like humans, the model initially learns only the distinction between legal (identical) and illegal (non-identical) consonant pairs; with more exposure, it starts memorizing the individual consonant pairs in the training set. An important difference from the behavioral data is that model appears to stop generalizing to unattested identical pairs earlier than human participants; in fact, it is unclear when (if ever) human participants stop generalizing – they show a similar

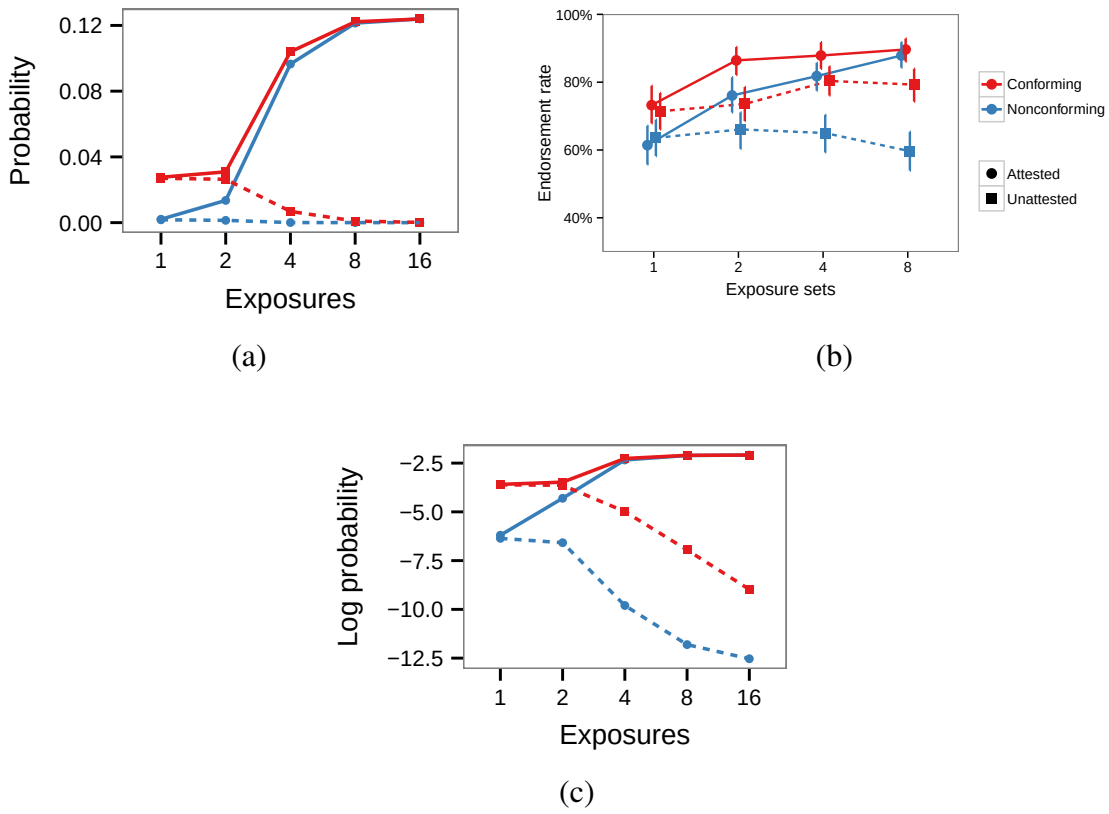


Figure 7.3: (a) Simulation of Experiment 2a: model predictive probability; (b) empiricial endorsement rates from Experiment 2a; (c) model log-transformed predictive probability.

willingness to generalize after 16 exposures as after 4 exposures (in absolute terms). When viewed on a log-scale, however (Figure 7.3c), it becomes clear that the model's predictive probabilities aren't identical between legal and illegal unattested segments – while both probabilities are low, the model still assigns higher probability to legal than to illegal segments.

Table 7.3 shows the most common clusters, weighted by the number of items in the data that were assigned to the cluster (as before). The number of members in each class sequence can be quite large; the class sequence ([], []), for example, contains more than 100 consonant pairs. We therefore show a sample of 8 class members when the class has more than 8 members.

After one exposure, the most common class sequence is ([], []), followed by ([], X); as expected after relatively little exposure, the learner assigns a fairly high probability to all segment sequences, though it prefers sequences of identical segments (without phonological restrictions on

Exposures	Class specification	Members	Probability
1	[], []	(k, s), **(θ, d)**, **(d, ʒ)**, **(f, k)**, **(g, k)**, **(n, z)**, **(r, b)**, **(t, ʃ)**	0.477
	[], X	(ʃ, ʃ), (g, g), **(ð, ð)**, **(θ, θ)**, **(f, f)**, **(l, l)**, **(t, t)**, **(v, v)**	0.421
	[], [-sonorant]	**(ʃ, z)**, **(ʒ, v)**, **(n, ð)**, **(p, v)**, **(r, ʒ)**, **(t, t)**, **(z, θ)**, **(z, b)**	0.005
	[-continuant], X	(g, g), (p, p), **(b, b)**, **(d, d)**, **(k, k)**, **(m, m)**, **(t, t)**	0.005
2	([], X)	(g, g), (n, n), **(ð, ð)**, **(θ, θ)**, **(d, d)**, **(m, m)**, **(s, s)**, **(z, z)**	0.413
	([], [])	**(θ, ð)**, **(ʒ, ʒ)**, **(b, v)**, **(f, z)**, **(m, n)**, **(r, d)**, **(v, b)**, **(v, n)**	0.245
	([-sonorant], X)	(ʃ, ʃ), (g, g), (p, p), **(θ, θ)**, **(ʒ, ʒ)**, **(d, d)**, **(k, k)**, **(t, t)**	0.007
	([-voice], X)	(ʃ, ʃ), (p, p), **(θ, θ)**, **(f, f)**, **(k, k)**, **(s, s)**, **(t, t)**	0.006
4	([-continuant, labial, -voice], X)	(p, p)	0.043
	([dorsal, +voice], X)	(g, g)	0.033
	([-anterior, -voice], X)	(ʃ, ʃ)	0.024
	([-distributed, +nasal], X)	(n, n)	0.018
8	([-continuant, labial, -voice], X)	(p, p)	0.059
	([dorsal, +voice], X)	(g, g)	0.035
	([-continuant, labial, -sonorant, -voice], X)	(p, p)	0.025
	([-continuant, -nasal, labial, -voice], X)	(p, p)	0.023

Table 7.3: Simulation of Experiment 2a: Classes with highest posterior probability. When the class sequence has more than 8 members, a random subset of 8 members are shown. The results of the 16-exposures condition are very similar to the 8-exposures one and are omitted for brevity.

the segment). Note that although the probabilities of ([], []) and ([], X) are similar, the probability of any individual non-identical pair derived from ([], []) is lower than the probability of any individual identical pair derived from ([], X), since there are many more pairs in ([], []) than in ([], X) and the particular pair generated by the model is sampled uniformly from the pairs in the class. The picture after two exposures is similar, though the probability of ([], X) becomes higher than the probability of ([], []).

After four and eight exposures, the classes with the highest probability only contain a single consonant pair – this is the model's way of memorizing individual pairs (though as before the identity of a particular pair is never represented in the model – the only representations are sequences of natural classes). The top four highest probability classes all include identical items only. This is an artifact of the fact that there are multiple feature matrices that pick out the same set of segments (e.g., [-sonorant, +voice] and [-nasal, -sonorant, +voice] both include all and only voiced obstruents – see also Table 7.2). This issue is compounded when there are two independent sets of segments (the first and second consonant) that can each be represented in multiple ways – the probability of each combination will be lower than the probability of sequences with a vacant slot that only include a single phonological class. Note that this doesn't affect the predictive distribution (Figure 7.3), since that distribution averages across all samples, regardless of the way each of the samples represents the parameter.

Finally, Figures 7.2c and 7.2d show the distribution of hyperparameters after one and eight exposures respectively. After one exposure, the number of clusters is typically two or three. The cluster parameters are typically sparsely specified (e.g., ([], [])), leading the model to conclude that the specification probability is low.

After eight exposures, the number of clusters is usually exactly eight (one for each consonant pair) – note that the consonant pairs in Experiment 2 were carefully chosen such that no subgeneralization across pairs are possible. This contrasts with Experiment 1, where the model was able to capture the attested segments precisely without positing a cluster for each individual segment, since there were subgeneralizations across attested segments that did not include any unattested

segments. The estimated specification probability in the 8 exposure condition is higher than in the 1 exposure condition, as expected. Note that in neither the 1-exposure nor the 8-exposure condition was the model able to precisely estimate the spreading probability q (the probability that the second class in the pair is a vacant slot), though there is somewhat less uncertainty about this parameter after 8 exposures.

7.5 Discussion and future work

7.5.1 Prior distribution on classes

Our model makes several uniformity assumptions:

1. All features are equally likely to be specified (all with probability p).
2. All feature values are equally likely; e.g., given that the *place* feature is specified, [labial], [coronal] and [dorsal] are all selected with probability $1/3$.
3. Sounds are generated uniformly from a class, e.g., the class [+voiced, labial, -sonorant] generates both [b] and [v] with probability $1/2$.

At least some of these assumptions are likely to be inconsistent with the priors that human participants bring into the experiment. For example, participants may believe that some features are more likely than others to participate in phonotactic patterns, either because of the articulatory or acoustic properties of the sounds or because of biases that are derived from the phonotactics of their native language (in our case, English). These priors can in principle be estimated from the lexicon of the participants' native language or from cross-linguistic typological tendencies.

It may be useful to explore the consequences of removing redundant features from feature matrices. For example, adding the feature [-nasal] to [-sonorant, +voice] does not change the set of the segments in the class. Removing redundant features could improve the interpretability of the results; moreover, it is likely to change the results themselves somewhat (perhaps in a desirable way).

7.5.2 Linking hypothesis

There is no well established linking hypothesis between probabilistic models of phonotactics and acceptability judgments. We did not attempt to propose one, and instead compared the model's predictions to the behavioral data in a qualitative way. Previous studies have either used a best-fit power function to relate probabilities to ratings, without any principled cognitive basis (Albright, 2009; Hayes & Wilson, 2008), or simply used raw correlations (Daland et al., 2011). In general, it is unclear whether subjective probability estimates are the only determinant of acceptability judgments in any linguistic domain (A. Clark, Giorgolo, & Lappin, 2013); finding an appropriate linking function remains a question for future research.

7.5.3 Sustained generalization

Humans were more willing to generalize to unattested legal segments or consonant pairs than the model's probability estimates predicted. There are several potential reasons for this finding. First, it is possible that humans do not extract as much information from the training set as the model does, due to noisy perception or memory constraints; for example, humans may extract the same amount of information from 16 exposures as the model does from 4 exposures. Second, a different probability model may be needed that explicitly reserves more probability mass to unattested segments drawn from legal classes. Third, the linking hypothesis may need to be revised; Figure 7.3c shows that there are sustained difference between legal and illegal unattested segments when considered on a log-scale.

From the empirical perspective, it may be of interest to explore at what point (if any) human participants stop generalizing to unseen segments. In a sense it would be surprising (though interesting) if humans still generalized after hundreds of exposures to each segment; at the same time, English speakers do generalize from [sl] to [sɪ], even though there are dozens of words in the English lexicon that start with [sl] and no words that start with [sɪ].

7.5.4 Pattern extraction

Our model requires significant supervision: the speech stream needs to be segmented into words and onsets need to be extracted from the words. To make the model more realistic, it would need to incorporate a syllable parser that would extract onsets automatically (Coleman & Pierrehumbert, 1997; Daland et al., 2011). It may also be beneficial to combine it with a word segmentation model (Adriaans & Kager, 2010).

7.6 Conclusion

This chapter proposed a generative probabilistic model of phonotactic generalization, and showed how its generalization behavior depends on the amount of exposure it receives to the language. We compared the predictions of the model to human behavioral data. The findings were as follows:

1. Like humans, the model learns broad patterns earlier than narrow ones, and can generalize from a single example.

2. After more exposure to the language, the model is less willing than humans to generalize to legal items that were not part of the training set.

We suggested possible extensions to the model that may achieve a better fit to human behavior.

8 Conclusion

This thesis reported on a series of experimental and computational projects that explored the implications of concurrently considering multiple probabilistic hypotheses in language processing and learning. Part 1 investigated how processing during language comprehension is affected by the number and distribution of the representations that are active at a given point in time, as measured by their entropy. Three studies were reported. The first study (Chapter 3) found that entropy over parses affected reading times in the way predicted by the entropy reduction hypothesis; this was only the case when entropy was calculated over full parses of the sentence rather than over predictions about immediate syntactic structure. The second study (Chapter 4) investigated the effect of typical syntactic context on single word recognition using MEG, and found that higher entropy over syntactic contexts was associated with lower neural activity in the left anterior temporal lobe. Finally, the third study (Chpater 5) applied the ideas from information theory developed in earlier chapters to sequential segment processing in spoken word recognition, and found that neural activity in auditory cortex reflected the predictability of the segment being processed. A central goal of the study was to identify effects of morphological prediction over and above segment prediction; the evidence for such effects was inconclusive.

Part 2 turned to language learning. It presented experimental data showing that a central assumptions in some models of acquisition are incorrect. According to the Subset Principle, learners are conservative: they select the narrowest generalization compatible with the learning data. Consequently, a generalization is only formed when two (or more) of its specific instantiations are

learned. In reality, participants learned broad generalizations before narrow ones, and were able to generalize from a single example (Chapter 6). The pattern of findings of the experiment is generally consistent with the predictions of a Bayesian model proposed in Chapter 7. Rather than being conservative, the model learns patterns at multiple levels of generality at once. It aims to explain the learning data in the most concise way possible. In the beginning of learning, the model prefers a single, simple pattern to multiple complex patterns that capture the data exactly; as it receives more training data, it becomes clear that the general pattern isn't empirically adequate, and the model reverts to a less concise but more accurate set of specific patterns. This straightforward mechanism to retreat from overgeneralization obviates the need for conservative learning.

References

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., & Ochiai, M. (2001). Reduction of nonperiodical extramural magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Transactions on Applied Superconductivity*, *11*(1), 669–672.

Adelman, J., & Brown, G. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214.

Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*(3), 311–331.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41.

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 58–69). Stroudsburg, PA: Association for Computational Linguistics.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Allen, K., Pereira, F., Botvinick, M., & Goldberg, A. (2012). Distinguishing grammatical constructions with fMRI pattern analysis. *Brain and Language*, *123*(3), 174—182.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560.

Astheimer, L., & Sanders, L. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia*, *49*(12), 3512–3516.

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. New York: Henry Holt.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313.

Baayen, R. H., Lieber, R., & Schreuder, R. (1997). The morphological complexity of simplex nouns. *Linguistics*, *35*(5), 861–878.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Milin, P., Djurdjević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Baayen, R. H., & Piepenbrock, R. (1995). *The CELEX Lexical Database (Release 2) [CD-ROM].* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex

words: A regression study across tasks and modalities. *The Mental Lexicon*, *2*(3), 419–463.

Balling, L. W., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, *23*(7-8), 1159–1190.

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*, 80–106.

Balota, D., Yap, M., & Cortese, M. (2006). Visual word recognition: The journey from features to meaning (a travel update). In *Handbook of psycholinguistics* (Vol. 2, pp. 285–375). Amsterdam: Academic Press.

Balota, D., Yap, M., Hutchison, K., Cortese, M., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Bar-Hillel, Y., Perles, M., & Shamir, E. (1964). On formal properties of simple phrase structure grammars. *STUF-Language Typology and Universals*, *14*(1-4), 143–172.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., & Bolker, B. (2014). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 0.999999-0)

Becker, M., & Levine, J. (2010). Experigen: An online experiment platform. *Available (April 2013) at https://github. com/tlozoot/experigen*.

Bemis, D., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, *31*(8), 2801–2814.

Bemis, D., & Pylkkänen, L. (2013). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, *23*(8), 1859–1873.

Berent, I., Marcus, G., Shimron, J., & Gafos, A. (2002). The scope of linguistic generalizations: Evidence from Hebrew word formation. *Cognition*, *83*(2), 113–139.

Berent, I., Shimron, J., & Vaknin, V. (2001). Phonological constraints on reading: Evidence from the Obligatory Contour Principle. *Journal of Memory and Language*, *44*(4), 644–665.

Berent, I., Wilson, C., Marcus, G., & Bemis, D. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry*, *43*(1), 97–119.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, *24*(1), 57–65.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

Bi, Y., Wei, T., Wu, C., Han, Z., Jiang, T., & Caramazza, A. (2011). The role of the left anterior temporal lobe in language processing revisited: Evidence from an individual with atl resection. *Cortex*, *47*(5), 575–587.

Bicknell, K., & Levy, R. (2010). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Blachman, N. (1968). The amount of information that y gives about x. *IEEE Transactions on Information Theory*, *14*(1), 27–31.

Blazej, L., & Cohen-Goldberg, A. (2014). Can we hear morphological complexity before words are complex? *Journal of experimental psychology. Human perception and performance*.

Blevins, J. P. (2013). The information-theoretic turn. *Psihologija*, *46*(4), 355–375.

Boguraev, B., & Briscoe, T. (1987). Large lexicons for natural language processing: utilising the grammar coding system of LDOCE. *Computational Linguistics*, *13*(3-4), 203–218.

Boland, J. (2005). Visual arguments. *Cognition*, *95*(3), 237–274.

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors

of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brennan, J., Lignos, C., Embick, D., & Roberts, T. P. (2014). Spectro-temporal correlates of lexical access during auditory lexical decision. *Brain and Language*, *133*, 39–46.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*, 163-173.

Brennan, J., & Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, *60*, 1139–1148.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Cappelle, B., Shtyrov, Y., & Pulvermüller, F. (2010). Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain and Language*, *115*(3), 189–201.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430.

Chen, Z., Hunter, T., Yun, J., & Hale, J. (2014). Modeling sentence processing difficulty with a conditional probability calculator. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Clark, A., Giorgolo, G., & Lappin, S. (2013). Statistical representation of grammaticality judgements: The limits of n-gram models. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 28–36). Sofia, Bulgaria:

Association for Computational Linguistics.

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. Malden, MA: Wiley-Blackwell.

Clark, R., & Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry*, 299–345.

Clifton, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, *2*(2), 234–250.

Colavin, R., Levy, R., & Rose, S. (2010). Modeling OCP-Place in Amharic with the Maximum Entropy phonotactic learner. In *Proceedings of the 46th meeting of the Chicago Linguistics Society.*

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017*.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, *30*(2), 234–250.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, *4*(2), 259–285.

Cristia, A., & Peperkamp, S. (2012). Generalizing without encoding specifics: Infants infer phonotactic patterns on sound classes. In A. K. Biller, E. Y. Chung, & A. E. Kimball (Eds.), *Proceedings of the 36th Annual Boston University Conference on Language Development (BUCLD 36)* (pp. 126–138). Somerville, MA: Cascadilla Press.

Cristià, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development*, *4*(3), 203–227.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Dahan, D. (2010). The time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, *19*(2), 121–126.

Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234.

Dale, A., Liu, A., Fischl, B., Buckner, R., Belliveau, J., Lewine, J., & Halgren, E. (2000). Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67.

Dale, A., & Sereno, M. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, *5*(2), 162–176.

Davis, M., Marslen-Wilson, W., & Gaskell, M. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218–244.

Dell, F. (1981). On the learnability of optional phonological rules. *Linguistic Inquiry*, *12*(1), 31–37.

DeLong, K., Urbach, T., Groppe, D., & Kutas, M. (2011). Overlapping dual erp responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207.

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

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Desikan, R., Ségonne, F., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., ... others (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*(1), 55–64.

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, *110*(3), 293–321.

Di Sciullo, A.-M., & Williams, E. (1987). *On the definition of word*. Cambridge, MA: MIT Press.

Dronkers, N., Wilkins, D., Van Valin, R., et al. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1-2), 145–177.

Duffy, S., Morris, R., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446.

Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L., Hare, M., & McRae, K. (2005). Cues, constraints, and competition in sentence processing. In *Beyond nature-nurture: Essays in honor of Elizabeth Bates* (pp. 111–138). Mahwah, NJ: Lawrence Erlbaum Associates.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Ettinger, A., Linzen, T., & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and Language*, *129*, 14–23.

Farnetani, E., & Recasens, D. (2013). Coarticulation and connected speech processes. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences, second edition* (pp. 316–352). Malden, MA: Wiley Online Library.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS ONE*, *8*(10), e77661.

Ford, M., Davis, M., & Marslen-Wilson, W. (2010). Derivational morphology and base morpheme

frequency. *Journal of Memory and Language*, *63*(1), 117–130.

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

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

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, *54*(3), 287–295.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, *77*, 91–106.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Fruchter, J., Linzen, T., Westerlund, M., & Marantz, A. (2015). Lexical preactivation in basic linguistic phrases. *Journal of Cognitive Neuroscience*. Advance online publication. doi: 10.1162/jocn_a_00822.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research*

Methods, *36*(3), 432–443.

Gallagher, G. (2013). Learning the identity effect as an artificial language: Bias and generalisation. *Phonology*, *30*(2), 253–295.

Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80–89.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Goldwater, S., Johnson, M., & Griffiths, T. L. (2006). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 459–466). Cambridge, MA: MIT Press.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*(4), 403–434.

Gouskova, M., Newlin-Łukowicz, L., & Kasyanenko, S. (2015). Selectional restrictions as phonotactics over sublexicons. *Lingua*. (Retrieved from http://www.nyu.edu/projects/gouskova/downloads/gouskova_newlin-lukowicz_kasyanenko_2015.pdf)

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460.

Green, M. J., & Mitchell, D. C. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *55*(1), 1–17.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, *5*, 162–181.

Grenander, U. (1967). *Syntax-controlled probabilities* (Tech. Rep.). Division of Applied Mathematics, Brown University.

Grishman, R., Macleod, C., & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th Conference on Computational Linguistics* (pp. 268–272).

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194–205.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Pittsburgh, PA: Association for Computational Linguistics.

Hale, J. (2003a). *Grammar, uncertainty and sentence processing* (Unpublished doctoral dissertation). Johns Hopkins University.

Hale, J. (2003b). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672.

Hale, M., & Reiss, C. (2003). The Subset Principle in phonology: Why the tabula can't be rasa. *Journal of Linguistics*, *39*(2), 219–244.

Halgren, E., Dhond, R., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J., ... others (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, *17*(3), 1101–1116.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The view from Building 20* (pp. 111–176). Cambridge, MA: MIT Press.

Halle, M., & Marantz, A. (1994). Some key features of distributed morphology. *MIT working papers in linguistics*, *21*(275), 88.

Hämäläinen, M., Hari, R., Ilmoniemi, R., Knuutila, J., & Lounasmaa, O. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*(2), 413–497.

Hare, M., McRae, K., & Elman, J. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, *48*(2), 281–303.

Hayes, B. (2011). *Introductory phonology*. Malden, MA and Oxford: Wiley-Blackwell.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, *41*(6), 401.

Humphries, C., Binder, J., Medler, D., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Jelinek, F., & Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, *17*(3), 315–323.

Jennings, F., Randall, B., & Tyler, L. (1997). Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes*, *12*(4), 485–504.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*,

24(4), 613–632.

Juhasz, B. J., & Berkowitz, R. N. (2011). Effects of morphological families on English compound word recognition: A multitask investigation. *Language and Cognitive Processes*, *26*(4-6), 653–682.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing*. Pearson Education India.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

Kallmeyer, L. (2010). *Parsing beyond context-free grammars*. Heidelberg: Springer.

Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. *LiLT (Linguistic Issues in Language Technology)*, *11*.

Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1-2), 43–73.

Kennison, S. (2001). Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin & Review*, *8*(1), 132–138.

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

Korhonen, A., Krymolowski, Y., & Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th international conference on language resources and evaluation.* Genova, Italy.

Kostić, A. (1991). Informational approach to the processing of inflected morphology: Standard data reconsidered. *Psychological Research*, *53*(1), 62–70.

Kostić, A. (1995). Information load constraints on processing inflected morphology. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 317–344). Hillsdale, NJ: Erlbaum.

Kostić, A., Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. Baayen & S. R. (Eds.), *Morphological structure in language processing* (pp. 1–43). Berlin: Mouton de Gruyter.

Kucera, H., & Francis, W. N. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*(7-8), 1089–1132.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*, 190–207.

Kutas, M., & Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Stroudsburg, PA: Association for Computational Linguistics.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 1055–1065). Stroudsburg, PA.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). London and New York: Psychology Press.

Levy, R., & Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*, 229.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lin, Y., & Mielke, J. (2008). Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics*, *14*(1), 214–254.

Linzen, T., & Gallagher, G. (2014). The timecourse of generalization in phonotactic learning. In J. Kingston, C. Moore-Cantwell, J. Pater, & R. Staub (Eds.), *Proceedings of Phonology 2013*. Washington, DC: Linguistic Society of America.

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

Linzen, T., Marantz, A., & Pylkkänen, L. (2013). Syntactic context effects in visual word recognition: An MEG study. *The Mental Lexicon*, *8*(2), 117–139.

Linzen, T., & O'Donnell, T. J. (2015). A model of rapid phonotactic generalization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP) 2015*.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge, UK and New York: Cambridge University Press.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, M. P. (1980). *A theory of syntactic recognition for natural languages*. Cambridge, MA, USA: MIT Press.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG and MEG data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*(1), 3–33.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Mazoyer, B., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., ... Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479.

McCarthy, J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207–263.

McDonald, S., & Shillcock, R. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652.

McDonald, S., & Shillcock, R. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295–323.

McMahon, A. M. (2002). *An introduction to English phonology*. Edinburgh: Edinburgh University Press.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Melinger, A., & Dobel, C. (2005). Lexically-driven syntactic priming. *Cognition*, *98*(1), B11–B20.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 1245994.

Meunier, F., & Segui, J. (1999). Frequency effects in auditory word recognition: The case of

suffixed words. *Journal of Memory and Language*, *41*(3), 327–344.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford and New York: Oxford University Press.

Milin, P., Filipović Djurdjević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*(1), 50–64.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. Blevins (Ed.), *Analogy in grammar: Form and acquisition* (pp. 214–252). Oxford: Oxford University Press.

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127.

Moreton, E. (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*, 165–183.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. H. (2005). Changing places: A cross-language perspective on frequency and family size in Dutch and Hebrew. *Journal of Memory and Language*, *53*(4), 496–512.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*(3), 721.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 705–741.

Nederhof, M.-J., & Satta, G. (2003). Probabilistic parsing as intersection. In *8th International Workshop on Parsing Technologies* (pp. 137–148). Nancy, France.

Norris, D. (1982). Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. *Cognition*, *11*(1), 97–101.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Oldfield, R. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Pallier, C., Devauchelle, A., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527.

Pater, J., & Moreton, E. (2012). Structurally biased phonology: Complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad*, *3*(2), 1–41.

Patterson, K., Nestor, P., & Rogers, T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Petrov, S., & Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 404–411). Stroudsburg, PA: Association for Computational Linguistics.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pulvermüller, F., Cappelle, B., & Shtyrov, Y. (2013). Brain basis of meaning, words, constructions, and grammar. In *Oxford handbook of construction grammar* (pp. 396–416). Oxford University Press.

Pulvermüller, F., Shtyrov, Y., Hasting, A. S., & Carlyon, R. P. (2008). Syntax as a reflex: Neurophysiological evidence for early automaticity of grammatical processing. *Brain and Language*, *104*(3), 244–253.

Pylkkänen, L., Feintuch, S., Hopkins, E., & Marantz, A. (2004). Neural correlates of the effects of morphological family frequency and family size: An MEG study. *Cognition*, *91*(3), B35–B45.

Pylkkänen, L., Llinás, R., & Murphy, G. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109.

Pylkkänen, L., & Marantz, A. (2003). Tracking the time course of word recognition with meg. *Trends in cognitive sciences*, *7*(5), 187–189.

Rao, R., Ballard, D., et al. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30–54.

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

Roark, B. (2011). Expected surprisal and entropy. *Oregon Health & Science University, Tech. Rep*.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333). Stroudsburg, PA: Association for Computational Linguistics.

Roark, B., & Johnson, M. (1999). Efficient probabilistic top-down and left-corner parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*

(pp. 421–428). Stroudsburg, PA: Association for Computational Linguistics.

Rodd, J., Davis, M., & Johnsrude, I. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266.

Rogers, T., Lambon Ralph, M., Garrard, P., Bozeat, S., McClelland, J., Hodges, J., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Roland, D., & Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. In P. Merlo & S. Stevenson (Eds.), *The lexical basis of sentence processing: Formal, computational, and experimental issues* (pp. 325–346). John Benjamins.

Rose, S., & King, L. (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech*, *50*(4), 451–504.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145–163.

Schmauder, A. (1991). Argument structure frames: A lexical complexity metric? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 49–65.

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2015). Competition between conceptual relations affects compound recognition: the role of entropy. *Psychonomic Bulletin & Review*, 1–15.

Scholes, R. J. (1966). *Phonotactic grammaticality*. The Hague: Mouton.

Schreuder, R., & Baayen, R. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Schuler, W., AbdelRahman, S., Miller, T., & Schwartz, L. (2010). Broad-coverage parsing using

human-like memory constraints. *Computational Linguistics*, *36*(1), 1–30.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100–107.

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

Shapiro, L., Zurif, E., & Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. *Cognition*, *27*(3), 219–246.

Shetreet, E., Friedmann, N., & Hadar, U. (2010). The neural correlates of linguistic distinctions: unaccusative and unergative verbs. *Journal of Cognitive Neuroscience*, *22*(10), 2306–2315.

Shetreet, E., Palti, D., Friedmann, N., & Hadar, U. (2007). Cortical representation of verb processing in sentence comprehension: number of complements, subcategorization, and thematic frames. *Cerebral Cortex*, *17*(8), 1958–1969.

Simon, D., Lewis, G., & Marantz, A. (2012). Disambiguating form and lexical frequency effects in MEG responses using homonyms. *Language and Cognitive Processes*, *27*(2), 275–287.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, *32*(25), 8443–8453.

Solomyak, O., & Marantz, A. (2009a). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042–2057.

Solomyak, O., & Marantz, A. (2009b). Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. *Brain and Language*, *108*(3), 191–196.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, *21*(2), 165–201.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in*

Cognitive Sciences, *13*(9), 403–409.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814.

Thompson, C., Bonakdarpour, B., Fix, S., Blumenfeld, H., Parrish, T., Gitelman, D., & Mesulam, M. (2007). Neural correlates of verb argument structure processing. *Journal of Cognitive Neuroscience*, *19*(11), 1753–1767.

Todorovic, A., & de Lange, F. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *The Journal of Neuroscience*, *32*(39), 13389–13395.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528–553.

Valiant, P., & Valiant, G. (2013). Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems* (pp. 2157–2165).

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394–417.

Van Petten, C., & Luka, B. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.

Wacongne, C., Changeux, J., & Dehaene, S. (2012). A neuronal model of predictive coding

accounting for the mismatch negativity. *The Journal of Neuroscience*, *32*(11), 3665–3678.

Wilson, K., & Carroll, J. (1954). Applications of entropy measures to problems of sequential structure. In *Psycholinguistics: A survey of theory and research* (pp. 103–110). Bloomington and London: Indiana University Press.

Wilson, M., & Garnsey, S. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language*, *60*(3), 368–392.

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

Wurm, L., Ernestus, M., Schreuder, R., & Baayen, R. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and conditional root uniqueness points. *The Mental Lexicon*, *1*(1), 125–146.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yarkoni, T., Speer, N., Balota, D., McAvoy, M., & Zacks, J. (2008). Pictures of a thousand words: Investigating the neural mechanisms of reading with extremely rapid event-related fMRI. *NeuroImage*, *42*(2), 973–987.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics 2008* (Vol. 123, pp. 5687–5690). New York: Acoustical Society of America.

Yun, J., Chen, Z., Hunter, T., & Hale, J. (2015). Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics*, *24*, 113–148.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.