Cognitive and communicative pressures in natural language

by

Kyle Mahowald

A.B., Harvard University (2009)
M.Phil., University of Oxford (2011)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .... Signature redacted
Department of Brain and Cognitive Sciences
May 6, 2016

Certified by....... Signature redacted ....
Edward Gibson
Professor
Thesis Supervisor

Accepted by . Signature redacted ....
Matthew Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

Cognitive and communicative pressures in natural language

by

Kyle Mahowald

Submitted to the Department of Brain and Cognitive Sciences
on May 6, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

Why do languages have the words they do instead of some other set of words? In the first part of this thesis, I argue that cognitive and communicative demands strongly influence the structure of the lexicons of natural languages. It is known that words in natural language are distributed such that shorter words are more frequent and occur after more predictive contexts. I provide evidence that, at least in part, this pattern is driven by word shortenings (i.e., chimp → chimpanzee) and that word shortenings can be predicted by principles of efficient communication. I also show that, using nonce words with no pre-existing semantic meaning, a Zipfian correlation between length and frequency emerges in freely produced text and that this correlation is driven by participants' tendency to *reuse* short words more readily than longer words. In addition to word length, I investigate phonetic probability in a corpus of 97 languages. Across a wide variety of languages and language families, phonetic forms are optimized for efficient communication. And, using baseline phonetic models, I show that the words in the lexicons of four languages (English, Dutch, German, and French) are more tightly clustered in phonetic space than would be suggested by chance alone.

This thesis depends on standard methods in language research. How reliable is the data that we work with as a field? In the second part of this thesis, I tackle that question by examining two dominant methods in modern language research: behavioral experiments (specifically syntactic priming) and linguistic acceptability judgments. I present data, based on large-scale surveys, showing that many of the standard syntactic and semantic judgments in a mainstream linguistic journal are flawed. Using this data, I construct a Bayesian prior over judgments and give recommendations for performing small sample-size experiments in linguistics that will not overly burden researchers. Finally, I present a large-scale meta-analysis of syntactic priming (the largest meta-analysis of a psycholinguistic phenomenon) and find that, while many priming studies are severaly underpowered, there is no evidence of intense p-hacking.

Thesis Supervisor: Edward Gibson
Title: Professor

Acknowledgments

I've occasionally looked at someone's thesis just to read the acknowledgments. If that's the reason you're here, then welcome! Sorry you're not sticking around.

The first person to thank here is my advisor Ted Gibson, followed closely by Ev Fedorenko. Ted and Ev are my friends as much as they are my advisors, and I couldn't have asked for better advisors as friends or friends as advisors. As friends, they have been there for me every step of the way: countless Flour lunches, trout/steak/burger/Ev salad dinners, retreats, trips, sporting events, Dixit, and so much more. As advisors, they helped me grow from someone with a background in English literature and syntax into a fully fledged quantitative psycholinguistics bunny who knows how to do all sorts of things. I have learned enormous amounts from both of them. While the advisorship half of the advisorship/friendship will necessarily change with my graduation, I look forward to the continuing flourishing of the friendship half.

There are many other mentors to thank, including the non-Tev part of my committee. Steve Piantadosi worked with me and advised me on much of the work in this thesis. His ability to solve seemingly every research problem and answer every question is a wonderful thing to have on one's side, and I am honored to have been among his first students. Josh Tenenbaum is one of the most exciting people there is to talk to about science; I left every conversation with him feeling energized. The same is true of Nancy Kanwisher, who always had something sharp and helpful to say every time I met with her and manages to do it in a way that leaves you convinced that she's the coolest.

I thank all of the collaborators on the work in this thesis – the aforementioned Ted and Ev and Steve, along with Anne Christophe, Isabelle Dautriche, Richard Futrell, Peter Graff, Jeremy Hartman, and Ariel James – as well as everyone who contributed to the individual pieces of research that went into this dissertation.

My fellow Tedlab graduate students and postdocs over the years have been enormously helpful in many ways: Leon Bergen, Peter Graff, Hal Tily, Tim O'Donnell, Alex Paunov, Richard Futrell, Melissa Kline, Laura Stearns, as have the many Evlabbies

including Olessia Jourvalev, Brianna Pritchett, Idan Blank, Zach Mineroff, and Caitlyn Hoeflin.

Among the many others I could list here, I thank my previous academic advisors: Daniel Donoghue at Harvard; Aditi Lahiri and Mary Dalrymple at Oxford. I am also very grateful to Bevil Conway, Vic Ferreira, Roger Levy, Josh McDermott, Alexa Khan, Julian Jara-Ettinger, and everyone else who made MIT a great place to work.

Next, I thank my shorter, darker-haired roommate of the last 6 years, Dan Roberts. It's been invaluable having such a good friend right in my apartment, even when that just means making time for Tevensons.

Summing up my family's contributions on this little page feels a bit inadequate. My mom and dad, my sister Megan, and Nan & Pop have supported my education every step of the way, from pre-K through this dissertation. They have given me every opportunity I ever could have asked for and much more.

And then there's ZOZ Robbie Kubala. I've now literally written a dissertation about words, but I still don't know any that can do him justice. So I'll just say that he's the best.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When Juliet says "A rose by any other name would smell as sweet," that may well be true from an olfactory perspective. But if a rose had some other name, would Juliet be able to access it just as easily from her mental library of stored words? If the Montagues and Capulets were having a noisy spat somewhere outside Juliet's balcony, would the ruckus cause Romeo to misunderstand Juliet and think that she said something about a nose smelling as sweet? Is it burdensome for Romeo and Juliet's language processing systems that, in addition to being a flower, rose can also be a woman's name and the past tense of rise? A rarer textual variant of Romeo and Juliet, Quarto 2, calls the phrase "any other name" by another name, stating that "a rose by any other word would smell as sweet." Would that version have become as memorable and widely cited?

The first part of this thesis is about an aspect of language that Juliet touches on when she muses about why a rose is called a rose: the relationship between word forms and their meanings. Imagine an alternative version of English in which I took all the meanings we might want to convey and randomly paired them with word forms. Would this bizarro version of English be just as efficient for communication? I explore that question in two distinct domains: first asking how the distribution of word lengths observed across languages, emerges through the communicative demands of speakers. Then, I turn to questions of word form: how the phonetic properties of words are affected by their use.

While there is no *a priori* reason why a *dog* should be called a *dog* and a *tarantula* a *tarantula* instead of the other way around, there is a good reason that the word *dog* is shorter: the word *dog* needs to be accessed much more frequently and therefore, from a communicative perspective, it is advantageous for it to be short. Indeed, across languages, lexicons have evolved such that most frequent words are short (Zipf, 1935) and predictable from context (Manin, 2006; Piantadosi, Tily, & Gibson, 2011). Besides the general property of lexicons that short words are predictable, people selectively use shorter words (Mahowald, Fedorenko, Piantadosi, & Gibson, 2012)–or omit words altogether–when the context is predictive (Frank & Jaeger, 2008; Levy & Jaeger, 2007).

Speakers also have degrees of freedom in the duration used to pronounce a given word. That is, a word can be pronounced differently based on its context (Lindblom, 1990), particularly its predictability in context (Gahl & Garnsey, 2004; Bell et al., 2003; Seyfarth, 2014), in a way that is sensitive not just to the speaker's needs but to the intended audience (Pate & Goldwater, 2015). Thus, we would expect that the word "dog" in the highly predictable sentence "I have to go home to walk my *dog*" would be pronounced more quickly than in the sentence "In the office I found a *dog*," where the italicized word could be just about anything. The logic is that, when the context is unpredictive, the listener needs more time to understand what's coming next and could benefit from a slower, more careful pronunciation. When the next word is known, the word can be pronounced quickly since there is little chance of misunderstanding it. These effects are consistent with the idea that linguistic utterances are structured so as to spread out the rate at which information is transmitted, in the sense of efficient coding theory (Shannon, 1948). By using shorter forms for more frequent and predictable meanings and longer forms for less frequent and less predictable meanings, languages establish a trade-off between the overall effort needed to produce words and the chances of successful transmission of a message (van Son & Pols, 2003; Aylett & Turk, 2004; Levy & Jaeger, 2007).

The efficient profile of word lengths shown across languages raises the question of how this efficient profile emerges, even as meanings of words change and the words themselves changes. In Chapter 2, I focus first on the question of how efficient word lengths evolve in the lexicon and how languages successfully navigate the tradeoff between the length of words and their robustness to noise. I test the hypothesis that, when a word is highly predictable in context, it will shorten in order to avoid being longer than it needs to be. To do that, I examine word shortenings like *chimpanzee* to *chimp* in a corpus analysis and a behavioral experiment (work published in Mahowald et al. (2012)). I then turn to a study investigating how, in the absence of meaning, speakers use novel words. Specifically, I give speakers 12 novel words (of varying length) and ask them to use some or all of those words in a science fiction story. In the stories produced, speakers show the same Zipfian distribution of word frequencies and the same length/frequency correlation seen in natural languages.

In Chapter 3, drawing on work in Mahowald, Dautriche, Piantadosi, and Gibson (Under revision), I focus on another prediction from Zipf: that the most frequent words should also be the easiest to produce and understand. That is, is there any systematic phonetic difference between frequent 3-phone words like *dog* and *cat* and less frequent 3-phone words like *nib* and *vim*? Specifically, we focus on *phonotactic probability* (which we approximate using orthographic probability) and *phonological neighborhood density.*

Phonotactic probability is a measure of the well-formedness of a string in a given language. For instance, in English the word *drop* is phonotactically quite probable, *dwop* is less probable but still allowed, and *dsop* has, essentially, zero probability. In this work, I will use orthographic probability, as measured by an n-gram model, as a proxy for phonotactic probability. Under a bigram model of orthographic probability, for instance, the probability of a string like *drop* would depend on the probability of the two-letter sequences that make up the word: *dr*, *ro*, and *op*. Phonological neighborhood density of a word w is the number of words that differ from word w by one insertion, deletion, or substitution (Luce, 1986; Vitevitch & Luce, 1998).

The result of Chapter 3–that the most frequent words in a language have the most neighbors and the highest phonotactic probability–raises the question whether phonological neighbors are desirable in a lexicon or whether they lead to increased competition and should be avoided. In Chapter 4, (work submitted as Mahowald, Dautriche, Piantadosi, Christophe, and Gibson (Under revision)), I evaluate the real lexicon against a simple model of phonotactics, asking whether phonological neighbors occur more often than one would expect based on chance alone.

This thesis presents work at the intersection of cognitive science and linguistics. In both of these areas, there has been recent controversy as to best scientific practice. The second part of the thesis focuses on questions in the methodology of linguistics and psycholinguistics. I ask whether dominant methods used in traditional linguistics (acceptability judgments in syntax and semantics) and psychology (behavioral experiments, using syntactic priming as a case study) produce robust, generalizable results.

In linguistics, especially research in syntax and semantics, a major source of data is acceptability judgments: judgments as to which sentences are acceptable as compared to others. In recent years, increases in the availability of data have led to discussions of the weaknesses of the informal method (Arppe & Järvikivi, 2007; Cowart, 1997; Gibson & Fedorenko, 2013, 2010a; Gibson, Piantadosi, & Fedorenko, 2013; Gross & Culbertson, 2011; Schutze, 2006; Sorace & Keller, 2005; Wasow & Arnold, 2005; Linzen & Oseki, 2015). Such weaknesses include potential cognitive biases on the part of the researcher and participants, difficulty in controlling for discourse context, the inability to find interactions among factors, and the inability to find probabilistic effects or relative effect sizes.

Gibson and Fedorenko (2010a), Linzen and Oseki (2015) and others have shown that this method leads to errors in the literature: some of the judgments that underpin major theoretical claims are simply not widely accepted by the linguistic community. In work adapted from Mahowald, Graff, Hartman, and Gibson (2015), I present a large-scale evaluation of the reliability of the linguistic literature by randomly sampling experiments from 10 years of the journal *Linguistic Inquiry.* I use that data both to evaluate the literature and to act as a prior for future linguistic judgments (Mahowald, Graff, et al., 2015). We call this paradigm SNAP Judgments (Small N Acceptability Paradigm for Linguistic Judgments).

In effect, SNAP Judgments recommends making theoretical linguistics more empirically oriented by increasing the amount of empirical data and running more experiments. But has the pscyholinguistic enterprise produced reliable and robust data? The state of the field in other areas of psychology and experimental science more generally has suggested we cannot assume that psycholinguistics is producing consistent and clear results. There has been work in other branches of psychology alleging widespread publication bias (Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014; Landy & Goodwin, 2014), low reproducibility (Open Science Collaboration, 2015), and low statistical power (Button et al., 2013b, 2013a), which can lead to inflated false-positive rates in the literature and unreliable results (Gelman & Carlin, 2014).

As a case study, in Chapter 6, I describe results from Mahowald, James, Futrell, and Gibson (2016), a systematic meta-analysis of syntactic priming in language production. Syntactic priming (Bock, 1986; V. S. Ferreira & Bock, 2006; Pickering & Ferreira, 2008) is a major phenomenon in psycholinguistics that has been central to the field for 30 years. In effect, it states that, if someone is primed with a syntactic structure X that freely varies with syntactic structure Y, a person is more likely to use X than she otherwise would have.

We report three results: a standard meta-analysis, an analysis of publication bias, and recommendations for sample size in future priming studies. We take effect size to be the log odds ratio of the proportion of target structure produced in the prime condition to the proportion of target structure produced in the no-prime condition. For 45 of the 73 papers in our sample, we obtained raw data from the authors and used it to derive estimates of effect size and standard error. From the remaining papers, we estimated the effect size and standard error using the published estimates. Along with effect sizes and their associated standard errors, we also collected information on several key manipulations that can potentially modulate the priming effect, including the construction used, lexical repetition, lag, and whether the priming is within or across languages. Using these variables, we estimated the average effect size of syntactic priming given various experimental conditions.

As a secondary analysis, I assessed the extent to which the set of papers in our study suffer from publication bias and low power. To assess publication bias and statistical power, I used p-curve, a tool developed for that purpose which works by analyzing the distribution of significant p-values in the literature (Simonsohn, Nelson, & Simmons, 2014a, 2014b). Using the raw data gathered from the study authors, I present a power analysis and give guidelines on how to run syntactic priming studies with sufficient statistical power.

Thus, this thesis has two main goals: first to understand pressures on the structure of the lexicon and second to improve the standard for best practices in work at the intersection of these two fields. The hope is that, by seeking to study natural language while simultaneously scrutinizing our own methodologies, we can make meaningful progress towards both goals.

Part I

Lexicon

Chapter 2

Lexicon and word lengths

Zipf famously showed that word length and frequency are inversely correlated: shorter words tend to be more frequent. The reason for this relationship, according to Zipf, is to maximize efficiency by using short words–which take less effort to produce–more often than long ones (Zipf, 1949, 1935). Zipf's ideas about language and efficiency in some ways anticipated information theory (Shannon, 1948), which provides a mathematical framework for formally characterizing the efficiency of communicative systems. However, the application of such ideas to natural language waned in the second half of the twentieth century, perhaps in response to the rise of generative linguistics, which has typically eschewed efficiency-based models of language. Chomsky, for example, has repeatedly argued against communication-theoretic models of language (Chomsky, 1975).

Recently, however, there has been renewed interest in studying language as a communication system. Piantadosi et al. (2011) used insights from information theory to build upon Zipf's claim about word length. As applied to language, the information conveyed by a word can be quantified by its *surprisal*, a measure of how unpredictable a word is given its context: $-logP(W = w|C = c)$ (Hale, 2001; Levy, 2008b). This notion captures the intuition that words that are completely predictable from context ($-logP(W = w|C = c)$) convey no ($log(1) = 0$) bits of information. For instance, in the phrase "to be or not to *be*," the final *be* is almost entirely predictable from the preceding context so would have surprisal approaching 0. Conversely, words that are highly unpredictable from context will have surprisal values that tend towards infinity as the probability of the word given its context approaches 0. The final word in "to be or not to *kumquat*" would have a surprisal that, while not infinite, would be high indeed.

In any communication system with variable-length communicative units (codes), it is most efficient to assign shorter codes to those elements that convey less information. In language, it is natural to take words as the code units, and study whether shorter words forms are indeed assigned to less informative meanings. Piantadosi et al. (2011) show that this prediction holds—that shorter words tend to convey less information (as measured by an idealized statistical model) than longer words. This result improves on Zipf's theory of frequency-driven word lengths by explicitly considering a word's typical predictability in linguistic context. Like an efficient variable-length code, language is organized such that low-information words—not just more frequent words—are shorter than high-information ones.

An important consequence of this organizational structure is that a lexicon in which word length is a function of information content allows speakers to approach a uniform rate of information conveyance. This *Uniform Information Density* (UID) allows speakers to maximize information conveyed without exceeding the production/perceptual channel capacity (Genzel & Charniak, 2003; Van Son & Pols, 2003; Aylett & Turk, 2004; T. Jaeger, 2006; Levy & Jaeger, 2007; Frank & Jaeger, 2008; T. Jaeger, 2010). Much previous work on UID has shown that information density can be manipulated by factors outside lexical content: syntactic variation such as *that* omission (Levy & Jaeger, 2007; T. Jaeger, 2010), phonetic reduction and lengthening (Aylett & Turk, 2004; Bell et al., 2003), and contraction of phrases like *you are* to *you're* (Frank & Jaeger, 2008). A correlation between information content and word length (Piantadosi et al., 2011), however, suggests the possibility that even content words—which are typically perceived as most fundamental to meaning and thus hardest for speakers to manipulate—can be used to control information rate.

Piantadosi et al.'s corpus results alone, however, are not sufficient to draw this conclusion. Because Piantadosi et al. do not attempt to control for meaning and syntactic category, the relationship between word length and information could arise from broad differences among syntactic classes of words. For instance, the effect could be driven by function words being shorter and less informative than content words, by a large-scale difference between nouns and verbs, or by other unforeseen regularities in the corpora.

In this chapter, we first evaluate whether the information/word length relationship holds for words of the same class, one would want to measure average information content while varying word length and holding meaning and syntactic category constant. Short/long word pairs, like *chimp/chimpanzee, math/mathematics*, and *exam/examination*, offer precisely such a controlled comparison by providing near-synonyms that vary in length. If the information-word length effect holds for words of the same class, shorter words in these pairs are predicted on average to convey less information. We test this prediction in a corpus study and a behavioral experiment.

In the second section of this chapter, we investigate where the observed relationship between word length and frequency comes from since word shortenings clearly cannot explain the whole effect. To do that, we elicited narrative texts from experimental participants, in which they used 12 nonce words that varied in length. Because these words do not have any *a priori* meaning, it was not obvious that they should follow a Zipfian frequency distribution or that they should show the length/frequency relationship seen in natural language. We tested these properties in the stories and found that, although short words and long words were equally likely to be used at first, the short words were more likely to be re-used. We use these results to explore a memory-based explanation of the Zipfian length/frequency effect.

2.1 Info/information theory

2.1.1 Introduction

In this experiment, we evaluate whether word shortenings (like like *chimp/chimpanzee, math/mathematics*) are use according to optimal information-theoretic compression.

In addition to determining whether content words can be used to manipulate information rate, there is another important implication of studying the information content of short/long word pairs. A systematic difference in expected surprisal between short and long forms would serve as evidence that the information/word length relationship constitutes part of a speaker's abstract linguistic knowledge and is not solely a product of long-term linguistic evolution. In other words, the most plausible explanation for a systematic difference in surprisal between nearly synonymous noun pairs differing in length would be that speakers are sensitive to the relationship between word length and predictability and thus actively choose word-forms that conform to that relationship during on-line production. For example, they tend to choose *exam* after predictive contexts and *examination* after non-predictive contexts. The absence of such a difference, however, would suggest that Piantadosi et al.'s effect does not constitute part of an individual speaker's knowledge of language. One might instead conclude that the effect arises from differences among classes of words or because of long-term pressure for linguistic efficiency that does not extend to the level of active speaker choice.

2.1.2 Materials

Word pairs of the form *exam/examination* for both the corpus and behavioral study were selected by generating a list of possible candidates using a combination of CELEX (R. Baayen, Piepenbrock, & Gulikers, 1995), Wordnet (Fellbaum, 1998), and (Marchand, 1966). Word pairs were selected to ensure that the short and long form of each pair could be used interchangeably.

Because the corpus does not distinguish between different meanings of identically spelled words, pairs like *ad/advertisement* were used only in the behavioral experiment since *ad* is not just an abbreviation for *advertisement* but is also used in Latin expressions like *ad infinitum*, *ad nauseum*, etc. Moreover, multi-word forms, like *United States*, were included in the behavioral experiment but not in the corpus analysis due to limitations of the corpus. Thus, the corpus materials are a subset of the behavioral materials. Every effort was made to include any and all pairs that meet the criteria above.

2.1.3 Experiment 1: Corpus study

In the corpus study, we first used the data from Piantadosi et al. (2011) (an unsmoothed three-gram model from the Google corpus) to obtain average surprisal estimates for 22 short/long word pairs. Using the corpus, surprisal for each word w was estimated by the equation:

$$-1/N\sum_{i=1}^{n} logP(W = w|C = c_i) \tag{2.1}$$

where c_i is the context of the ith occurrence of w and N is the total frequency of w. Because the three-gram model was shown to be the most reliable by Piantadosi et al., context c_i was estimated here as the two words preceding word w.

Replicating Piantadosi et al., the mean surprisal for long forms (9.21) was significantly higher than that for short forms (6.90) (p =.004 by Wilcoxon signed rank test). Of the 22 pairs, 18 showed higher average surprisal for the long form than for its shorter counterpart. A linear regression (with unscaled variables) modeling difference in log frequency between short and long forms as a predictor for difference in surprisal revealed that the effect held even when controlling for the fact that short words tend to be more frequent than long ones. Although there was indeed an expected significant effect of difference in log frequency on difference in surprisal (t=-4.67, p<.001), an intercept of 1.45 (t=2.76, p =0.01) indicated that, when there was no difference in frequency between the forms, the mean surprisal of long forms was 1.45 higher than that of short forms.

Figure 2-1 shows the difference in average surprisal between the long and short form of each word pair plotted against the log corpus count of the pair's short and long form combined (i.e., the frequency of the pair as a unit). For ease of reading, only the short form is listed on the plot. The key feature of the plot is that most pairs fall above the line drawn at x=0.

These results demonstrate that the long form of a word carries more information

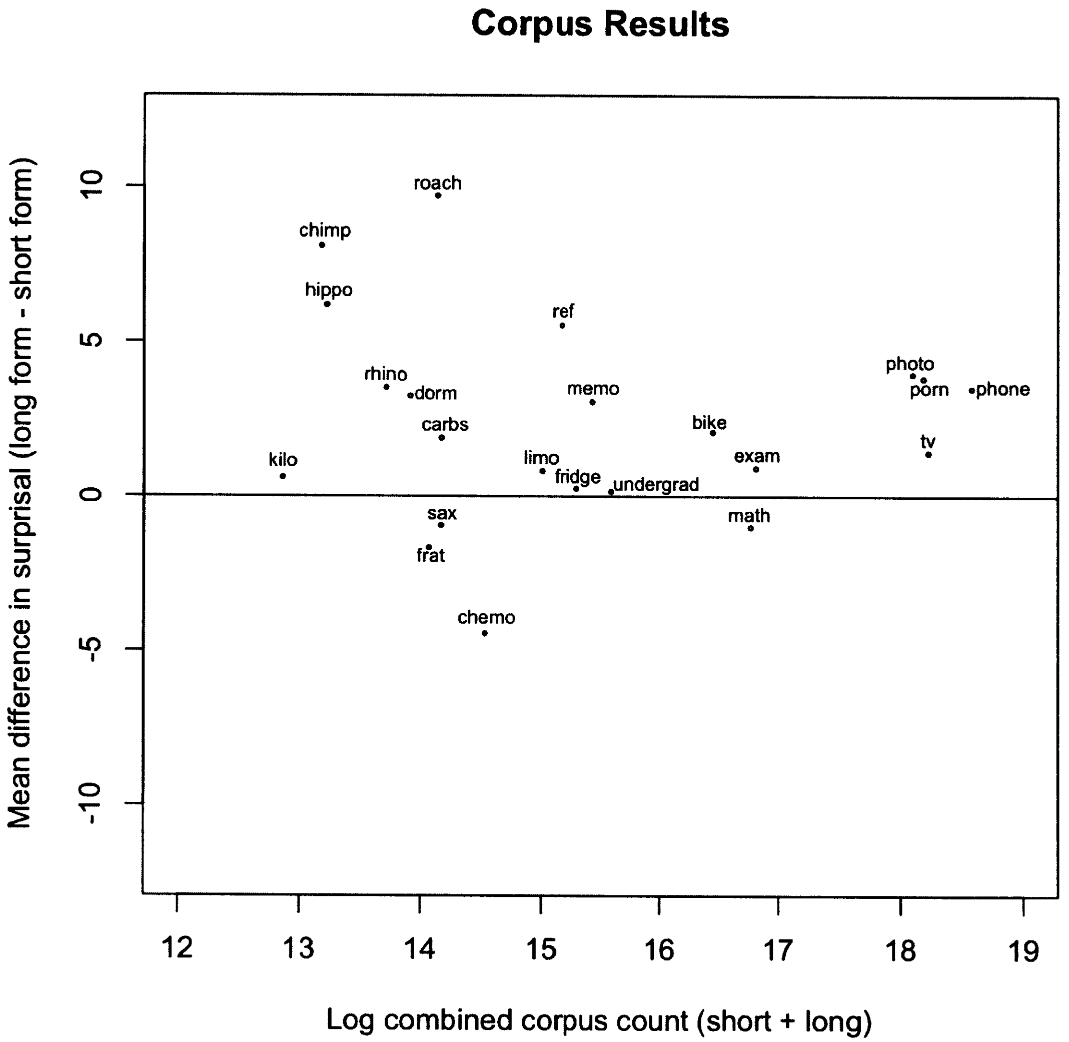


Figure 2-1: Difference in mean surprisal between the long and short form (long - short) plotted against log combined corpus count of short and long. The pairs above the line at x = 0 show the expected effect whereby long-form surprisal is greater than short-form surprisal.

on average than its shorter counterpart, and that this effect cannot be explained only by a difference in frequency between short and long forms: predictiveness of context plays an important role.

2.1.4 Experiment 2: Behavioral study

To test whether participants actively choose short forms in predictive contexts, we used Amazon's Mechanical Turk to present 58 native English speakers with forced-choice sentence completions in which they chose between the short and long form of a word pair based on which sounded more natural. The manipulation of interest was whether the context was predictive of the missing final word (supportive-context condition) or non-predictive (neutral-context condition).

Sample item:

Supportive context: Bob was very bad at algebra, so he hated...

1. math 2. mathematics

Neutral context: Bob introduced himself to me as someone who loved...

1. math 2. mathematics

The order of the answer choices was balanced across participants and items. Supportive and neutral contexts were matched for length. To avoid any biases from common phrases like "final exam," the key word was never presented as part of a common phrase. Comprehension questions were included to ensure that participants were engaged in the task.

To ensure that our context manipulation was effective, we presented a separate group of native English speakers (n=80) with the same sentence preambles and asked them to supply a word of their choosing to complete the sentence. For supportive contexts, the target word (either the short or long form) was chosen 52.4% of the time compared to just 1.6% of the time for neutral sentences. The effect of context on whether the word from the target pair was supplied was highly significant by a mixed-effect logistic regression with item and participant slopes and intercepts (β=5.22, z=13.09, P<10^{-15}).

In the critical experiment, in supportive contexts, the short form was chosen more often (67%) than in neutral contexts (56%). This effect of context on choice of form was significant when compared to the overall mean by a mixed-effect logistic regression (R. Baayen, Davidson, & Bates, 2008; Gelman & Hill, 2007), with item and participant slopes and intercepts (β=.75, z=3.65, P<.001). There was also a significant baseline preference for the short form independent of context (β=.77, z=2.76, P<.01). The context dependence of choice of form suggests that the correlation between word length and informativeness is likely influenced by language production phenomena, where users actively prefer to convey meanings with short forms when the meanings are contextually predictable, even when controlling for syntactic category and meaning.

Figure 2-2 shows the proportion of trials for which the long form of a word was selected in supportive contexts subtracted from the proportion of trials for which the long form was chosen in neutral contexts. As expected, the words tend to cluster above 0, which indicates that the long form of a word is chosen more often in neutral contexts.

We conclude that speakers actively select shorter word forms in more predictive contexts and longer forms in less predictive contexts.

2.1.5 Discussion

The information/word length effect holds for pairs of near synonyms that vary in length. The corpus analysis revealed that short forms in almost all instances have significantly lower information content than long forms. The word-choice experiment further showed that speakers are more likely to choose a short form in a supportive context than in a neutral context. We therefore conclude that, just as phonetic and syntactic factors can be used to manipulate information rate, so too can content words.

By looking at the same phenomenon (word shortening) in both a quantitative corpus analysis and a behavioral experiment with human participants, we have untied two disparate approaches to measuring linguistic context. Our corpus study relied on a precise notion of context: two words immediately preceding the target word. Although this measure of information is a rather impoverished substitute for the rich real-world context that determines a word's surprisal, it has the advantage of being precise and

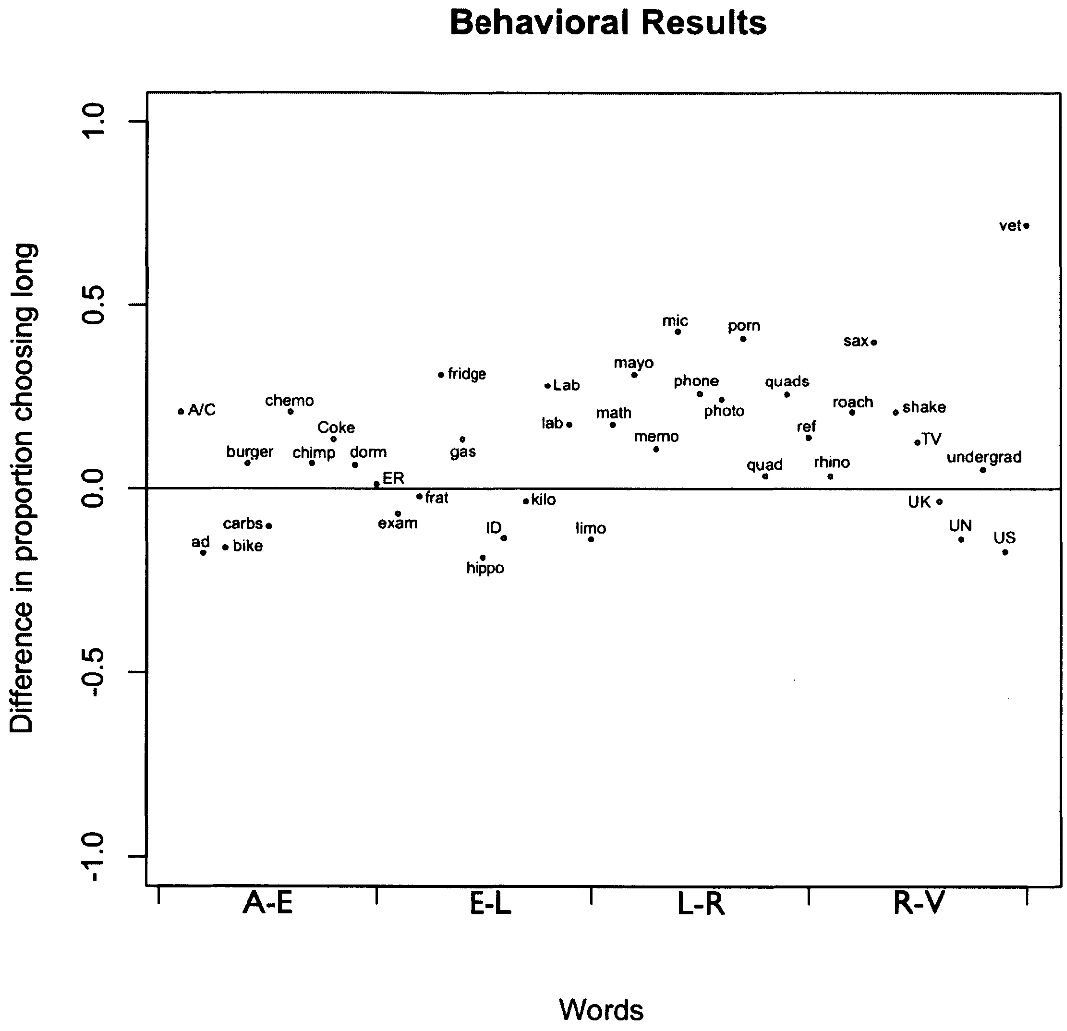


Figure 2-2: The y-axis shows the proportion of trials where the long form was chosen in predictive contexts subtracted from the proportion of trials for which the long form was chosen in non-predictive contexts. The pairs that fall above the line at x=0 show the expected effect whereby the long form was chosen more often in non-predictive contexts than in predictive contexts. The x-axis indicates the alphabetical order of the words.

easily quantified. Moreover, n-gram surprisal measures have repeatedly been shown to perform well in predicting many real-world linguistic phenomena like syllable duration (Aylett & Turk, 2004), phonological reduction (Jurafsky, Bell, Gregory, & Raymond, 2001), and word length (Piantadosi et al., 2011). N-grams also serve as an adequate surrogate for real-world context in a wide array of natural language engineering tasks (Jurafsky & Martin, 2009).

That said, n-gram measures of surprisal ignore the semantic and pragmatic cues that play a role in the interpretation of real-world linguistic context. Our behavioral study used an alternative notion of linguistic context in which a context c is deemed either "neutral" or "supportive" for a given word w based on whether speakers routinely supply word w after context c. Thus, for the behavioral study, whether or not a context was supportive was usually a product of salient semantic and pragmatic factors. The fact that our results generalized across both the precise n-gram notion of context as well as the fuzzier but more intuitive notion of context in the behavioral study is evidence that the relationship between word length and information content is not merely an artifact of using simplified n-gram-based measures of surprisal.

Moreover, these results suggest that considerations of word length and predictability form part of a speaker's knowledge of language. An important outstanding question concerns the level of abstraction at which this knowledge exists. It is possible that the corpus results presented here arise from learned preferences for each specific shortened form in predictive contexts and for each longer form in non-predictive contexts. Indeed, previous work (Fedzechkina, Jaeger, & Newport, 2012) has suggested that acquisition of communicatively optimal forms is easier than acquisition of less efficient forms. One way that this process could work would be for the short form to be preferentially learned in certain highly predictive contexts (*final exam*, *football ref*, etc.). But the behavioral experiment rules out this scenario *in general* since the sentences avoided these types of fixed phrases and presented participants with what were likely novel contexts. Thus, the behavioral results suggest that the corpus results (both the ones presented here and the ones reported by Piantadosi et al., 2011) arise because of an abstract association between word length and information. Such association is most plausibly due to speakers preferentially varying word length on-line during production. These results thus add to a growing body of work showing that speakers actively control information rate.

This line of research may also have implications for understanding certain types of lexical change. Specifically, if a word's surprisal decreases, one should expect that word to shorten over time. This hypothesis accords with long-held intuitions about language change: "[Shortened forms] originate as terms of a special group, in the intimacy of a milieu where a hint is sufficient to indicate the whole" (Marchand, 1966). Information theory provides a way to formalize this intuition. In future work, it may be possible to model and even predict lexical change based on changes in surprisal.

More broadly, these results demonstrate the power of applying information theory to the study of language. An information-theoretic framework and model was critical in formulating the behavioral experiment, and correctly predicted its outcome. Our results therefore provide further evidence that language–and the cognitive systems that process it—result in part from pressures for efficient communicative design.

Here we have claimed that the frequency with which one wants to talk about a *chimpanzee* in the world will affect the probability of shortening. Do we see a preference for short words even in the absence of semantics? In the next chapter, we ask whether word shortenings occur even without semantic effects like the ones discussed here.

2.2 Zipfian distributions in the absence of semantic differences

In the previous section, we saw that people selectively shorten words when the context is predictive. This account of the distribution of word lengths is largely based on the semantics of the words. That is, as a word meaning becomes more prominent and a word becomes more predictable in context, the word will shorten. In that sense, the length of a word is determined by a word's meaning and the frequency with which speakers want to convey that meaning.

Because most words are not shortenings of other words, word shortenings cannot explain the full range of effects that we see in natural language. Here, we ask whether the length/frequency effect emerges in the absence of any semantics at all by considering how speakers use a set of nonce words that vary in length. If even these nonce words–which have no *a priori* semantic meaning and do not obviously pick out different items in the real world–show the predicted length/frequency relationship, that would be evidence that the relationship between word length and frequency is not just mediated through semantics but is a part of speakers' knowledge of language.

Using a paradigm developed in Piantadosi (2014), we elicited raw text (a science fiction story) from experimental participants in which they were supplied with and invited to use twelve nonce words. Piantadosi (2014) showed that the frequency distribution of these nonce words was Zipfian. We varied the length of the nonce words from 2 to 12 letters and explore the relationship between the frequency of the nonce words in the stories and their length.

Piantadosi (2014) showed that, even among these nonce words, the rank-frequency distribution is Zipfian. He uses this result to argue against a class of models, known as "random typing models," that have been proposed to explain the Zipfian distribution of word frequencies (Miller, 1957; Ferrer i Cancho & Moscoso del Prado Martín, 2011; Conrad & Mitzenmacher, 2004). Under this account, one can imagine a language being generated by randomly typing on a keyboard (with a new word being generated by a space key). The resulting word distribution would be Zipfian, but as Piantadosi points out, this is not a plausible account of natural language. In natural language, words are stored and re-used.

How does the length/frequency correlation emerge in real words, if not just by shortenings and if not by random typing? One default possibility is that the length distribution in natural language is largely determined by the relationship between form and meaning. That is, over time, we assign short forms to meanings such that the most frequently accessed meanings have short code lengths. Under this account, we would not expect the frequency of use of a nonce word to vary based on its length in this task.

The other possibility is that people prefer using short words and will preferentially use short words in this task. Given a length-based difference in how people use the nonce words, there are two further possibilities. Speakers could a) preferentially choose short words from the list of possibilities or b) prefer to *re-use* short words. The former possibility would suggest a global preference for short words–that, given a choice between a short and a long word, the speaker would rather use the short word. The latter possibility would be consistent with a memory-based account of the word length distribution, whereby short words are easier to remember and therefore more likely to be re-used.

2.2.1 Experiment 1

Participants

We recruited 80 participants on Amazon's Mechanical Turk. Participants were limited to self-identified native English speakers with IP addresses inside the United States who produced at least half the requested amount of text. After exclusions, 59 participants remained.

Materials

The prompt was as follows:

An alien space ship crashes in the Nevada desert. Twelve creatures emerge:

[list of words]

In at least 2000 words, describe what happens next.

Your submission must be at least 2,000 words long. Your writing must be original, coherent, and based on the prompt.

The nonce words varied in length from 2 to 12 and are comprised of CV syllables. So a 4-letter word is CVCV (e.g., *bako*), a 12-letter word is CVCVCVCVCVCV (e.g., *bamoparatofa).* Each participant saw 2 words of each length for 12 words total.

Cleaning data

In order to asses the frequencies of each nonce word, we cleaned and spell-checked the data as follows. Plurals were counted as uses of the target word. Words were spell checked by looking for all words that are not in an English dictionary or in the stimuli set. Excluding 2-letter words (for which spelling errors are generally impossible to recognize, even by hand), we checked to see if there are 1-edit neighbors of a stimuli word. If so, we counted this as a use of the target word. Then, for words of length 5 and above that fit that criteria, we checked for 2-edit neighbors. We then looked at the corrections by hand and exclude inappropriate corrections (i.e., corrections that were clearly actually intended to be words).

Results

First, we examined whether the distrubion of nonce words used was Zipfian by computing, for each participant, the frequency and frequency rank of reach word. We found that, on a log-log plot of rank frequency against frequency, the words were roughly linear. This held even within words of the same length. That is, a given participant was more likely to use his or her most common 2-letter nonce word more often than his or her second most common nonce word in a way that was consistent with a Zipfian distribution. Figure 2-3 shows the rank frequency distributions.

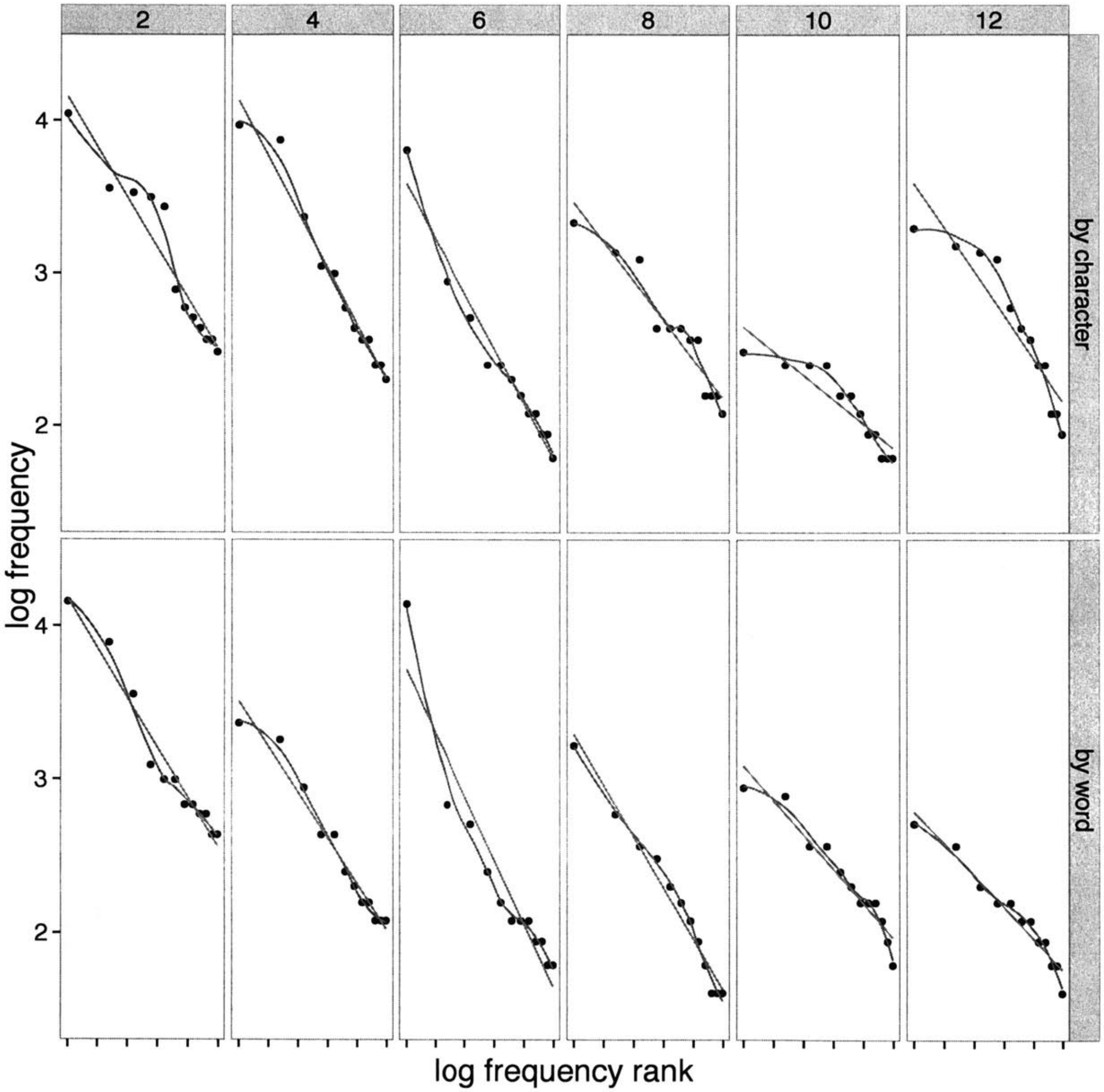


Figure 2-3: For Experiments 1 and 2, the log rank plotted against the log frequency for nonce words. Both experiments show Zipfian distributions.

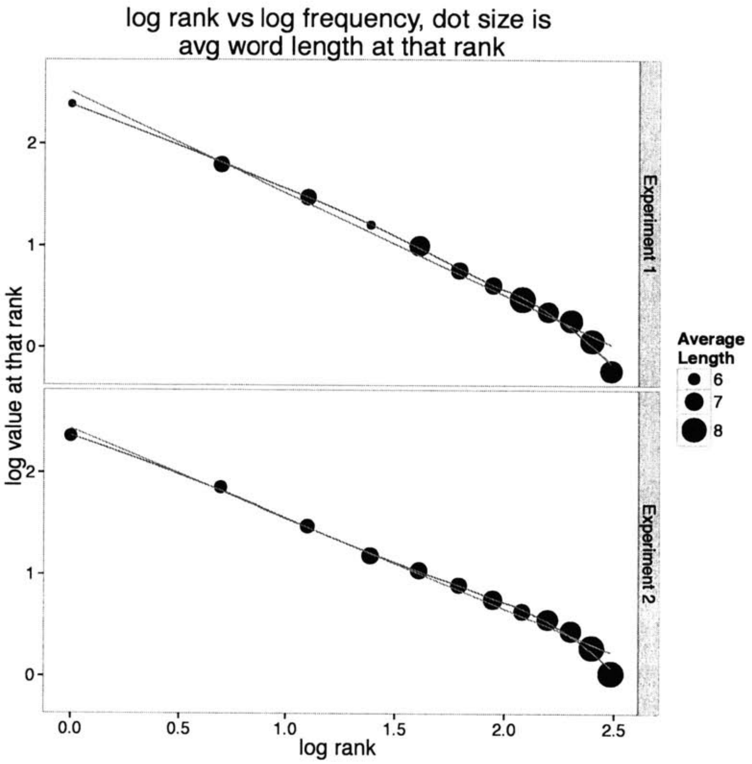


Figure 2-4: The log rank against the log frequency at that rank and shows the classic Zipfian shape. The size of the point is the average length of the word at that rank frequency. Note that the least frequent words are, on average, the longest.

We next tested for the presence of the length/frequency effect. We ran a Poisson regression to predict count from log length with random effect for worker (intercept and a slope by log length) and a random intercept for word. We found a sigificant effect of length such that a short word was more likely to be used than a long word (β=-.45, z = -4.4, p < .0001). Figure 2-4 plots the log rank against the log frequency at that rank and shows the classic Zipfian shape. The size of the point is the average length of the word at that rank frequency. Note that the least frequent words are, on average, the longest.

To assess where the length/frequency effect comes from, we examined how likely a word was to be used for the first time as a function of its length. Among words used, there is a trend (not significant) for short words to be used earlier in the story than later in the story (β=.012, t=1.78, p = .08), but the overall size of the effect is small. On the other hand, the probability of a short word being reused is much higher than

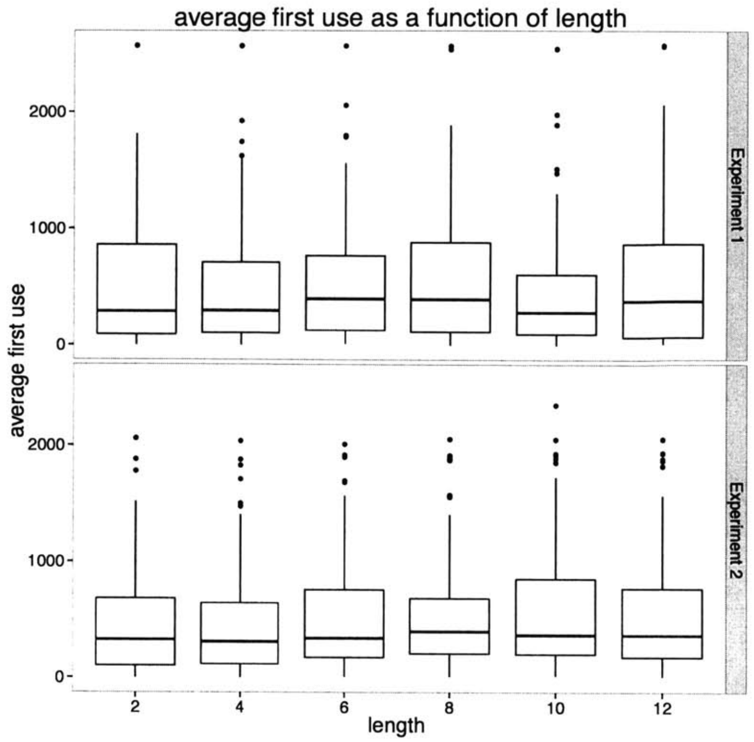


Figure 2-5: Separated by length, the average word position (from 0 to word n) of the first use of a particular word.

the probability of a long word being reused. In a regresssion predicting probability of re-use from length (with random effects of participant and word), there was a clear effect of length (β=-.15, z = -.51, p < .0001).

Thus, the effect for short words to be used more often is largely due to their being more likely to be re-used more often than longer words.

2.2.2 Experiment 2

There is a potential confound in Experiment 1 in that participants were asked to type 2,000 words of text. If they were interested in doing so as quickly as possible (while still receiving payment), they may have been incentivized to use short words as often as possible. To avoid this confound and to test whether the effects reported above replicate in a new sample, we ran a new version of the experiment with identical methods except that the story was required to be 10,000 characters instead of 2,000 words.

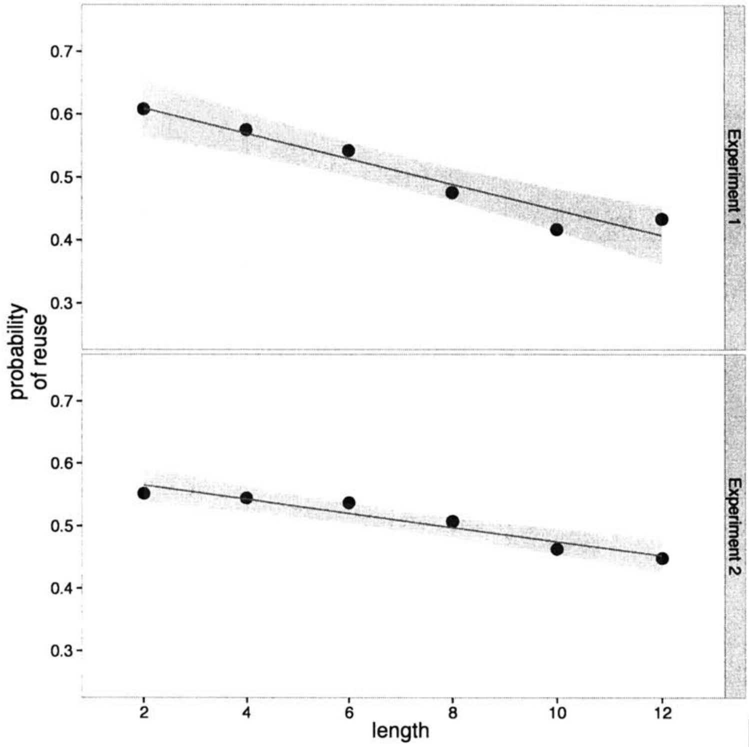


Figure 2-6: Separated by length, the average probability of re-using a word of that length, given that it was used once.

Participants

We recruited 80 participants on Amazon's Mechanical Turk. Participants were limited to self-identified native English speakers with IP addresses inside the United States who produced at least half the requested amount of text. After exclusions, 68 participants remained in this sample.

Materials

Materials were identical to Experiment 1 with the exception that we asked for 10,000 characters instead of 2,000 words.

Cleaning data

The cleaning procedure was identical to that in Experiment 1.

Results

We found similar effects across all measures to the effects observed in Experiment 1. There was a significant effect of log length on log frequency (β = -.27, z = -3.15, $p <$ .01). While there was a significant trend for a short word to be used before a long word (β=.025, t = 2.88, $p <$ .01), the effect is very small. Meanwhile, the effect of re-use is relatively strong such that short words are more likely to be re-used than long words after having been used once (β= - .090, z = -3.20, $p <$ .001).

Thus, as with Experiment 1, we found a Zipfian relationship between word length and word frequency. And, as with Experiment 1, the effect appeared to be driven by re-use and not initial use.

2.2.3 Discussion

Across two experiments investigating how speakers use nonce words in stories, we showed that a) a Zipfian rank/frequency distribution quickly emerges in the text samples, and b) a Zipfian relationship between word length and frequency emerges. The relationship between word length and frequency is of a similar magnitude to that observed in natural languages. Moreover, the correlation appeared to be largely driven by re-use and not by initial use of the words. That is, relative to a long word, a short word was only slightly more likely to be used as the first nonce word in a story. But it was *much* more likely to be re-used.

Crucially, by studying statistical properties of the distribution of these nonce words in this restricted environment, we can gain insights not available by studying naturalistic text alone. Natural language is, in a sense, confounded with the meanings that people want to convey. By removing any notion of semantics, we can probe what conditions are needed to observe the statistical properties we see in the languages of the world. Specifically, the results here are consistent with a usage-based account of the length/frequency distribution whereby speakers preferentially re-use short words because they are easier to remember, produce, and comprehend.

Critically, these findings (both in the experiments with word shortenings and in the present study) provide evidence against recent claims by Ferrer i Cancho and Moscoso del Prado Martín (2011), who argue that the observed correlation between word length and information content in lexical systems may not be meaningful, because such a correlation would also be found in random typing, which is not communicative. First, it should be noted that their observation cannot explain the primary finding of Piantadosi et al. (2011), that word length is better predicted from information content than from frequency since those two measures coincide for random typing models. Second, the mechanism proposed here–whereby short words are preferentially re-used after they have been used once–crucially relies on a stable notion of a word as a unit that is memorized and used.

More broadly, the experiments presented in this chapter suggest that, even in limited contexts, stable lexical characteristics emerge extremely quickly. In the word shortening study, we saw that speakers will show a preference for short words after predictitve contexts even in a very narrow forced choice task. And in the story generation study, we saw a system of word frequency distributions emerge in the minutes it took speakers to enter text.

Word length is one of the most well-studied quantitative features of natural language in part because it is easy to measure. But it is not the only feature of the lexicon that lends itself to predictions from information theory and the principles of efficient communicative design. In the next two chapters, we will investigate another such facet of the lexicon: the phonetic properties of words.

Chapter 3

Phonotactics and communication

3.1 Introduction

In this chapter, we are concerned not with length but in applying similar ideas to another way in which words can systematically vary: their phonetic form.

If a lexicon were structured for maximally efficient communication, how should word frequency be related to phonotactic probability and neighborhood density? Following the same logic used to explain why frequent words tend to be short, Zipf (1935) claimed that the Principle of Least Effort predicts that easily articulated sounds should be used more often in language than more difficult sounds. While Zipf was talking about individual sounds, there is compelling evidence that phonotactically probable words are easier to produce and understand in language use. For instance, the phonetics of languages evolve to enable easy articulation and perception (Lindblom, 1992, 1983, 1990) and the patterns of sounds observed across languages reflect articulatory constraints (Kawasaki & Ohala, 1980). Therefore, a language whose most frequent words are phonotactically *probable* likely requires less production effort than a language organized such that the most frequent strings are phonotactically *improbable*. I test this prediction in a Wikipedia corpus of 97 languages by examining the three-way correlations among frequency, phonotactic probability, and neighborhood density.

Thus, from a production perspective, it is more efficient for a language to be structured such that the most frequent words are phonotactically probable. But what about from the listener's perspective? Just as speakers have an easier time producing frequent sound sequences, listeners are also more adept at perceiving these sound sequences. Phonotactically probable words are more easily recognized than less probable words (Vitevitch, 1999). And there appears to be a learning advantage for probable strings: probable strings are learned more easily by infants and children (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013). These lines of evidence suggest a functional advantage to using phonotactically probable words not just from a production perspective but from a comprehension perspective as well.

Words that are phonotactically probable also tend to have many phonological neighbors. There are conflicting accounts as to how phonological neighborhood density affects language production and comprehension and therefore differing predictions about how the lexicon should be organized in terms of phonological neighborhood density. Under a simple noisy channel account, words with many neighbors cause increased processing difficulty since they are typically more confusable with other words. And indeed there is evidence that having many neighbors can have an inhibitory effect on lexical access in perception (Luce, 1986; Vitevitch & Luce, 1998) and elicit lexical competition that slows down word learning in toddlers (Swingley & Aslin, 2007). Moreover, Magnuson, Dixon, Tanenhaus, and Aslin (2007) shows that high-density word onsets inhibit reading times. We deal with these tradeoffs between sparsity and clumpiness more fully in Chapter 4.

To evaluate the extent to which the phonological forms of words may be explained by word usage, we investigated whether (a) wordforms that are orthographically probable (as measured over word types) are likely to be more frequent (by token) than wordforms that are less orthographically probable and (b) whether wordforms that are orthographically similar to other words are likely to be more frequent than phonologically more unique wordforms. In all of these analyses, we compared only words of the same length so that any resulting effects are not driven by word length.

By looking at orthographic probability and neighborhood density separately, we can attempt to discern whether the possible inhibitory affect of neighborhood density on language comprehension is reflected in the lexicon. If we observe a consistently *positive* correlation between frequency and orthographic probability and between frequency and phonological density, it would indicate that lexicons are structured so that the most commonly used words are easy to produce. In contrast, a consistently *negative* correlation would suggest that an overriding concern in lexical structure is that the most frequent words not be overly confusable with one another. If we observe no consistent correlation, it would suggest either that languages differ in how they have evolved or that there is no clear optimization in this regard. If we observe an effect for just phonotactics but a negative correlation for neighborhood density, it would suggest that there is a pressure for the most frequent words to be phonotactically probable but not confusable with other words.

These sorts of correlations have been examined in the literature before, but only for a small number of languages. Landauer and Streeter (1973) performed a similar analysis for English, and Frauenfelder, Baayen, and Hellwig (1993) for English and Dutch. All found that the most frequent words in the language have higher phonotactic probability and more phonological neighbors than more infrequent words. While these results are suggestive, it is difficult to draw conclusions based on a small set of related languages. It is particularly difficult to do so given the large differences that exist across language families in basic features of phonology like the number of distinct phonemes.

In the current study, we used orthographic lexicons from 97 typologically diverse languages downloaded from Wikipedia in order to investigate whether the relationship between orthographic probability, neighborhood density, and frequency reflect functional constraints. We found that frequent wordforms tend to be orthographically likely and have more neighbors than less frequent wordforms, suggesting that there is a functional pressure associated with word usage for languages to prefer phonotactically probable strings that are also phonologically more similar to one another.

West Germanic: Afrikaans, German, English, Luxembourgish, Low Saxon, Dutch, Scots, Yiddish, Alemannic; **Goidelic**: Irish, Scottish Gaelic; **Brythonic**: Breton, Welsh; **Hellenic**: Greek; **South Slavic**: Bulgarian, Macedonian, Serbo-Croatian, Slovene; **Albanian**: Albanian; **Iranian**: Central Kurdish, Persian, Kurdish, Mazandarani, Tajik; **Romance**: Aragonese, Asturian, Catalan, Spanish, French, Galician, Italian, Lombard, Neapolitan, Occitan, Piedmontese, Portuguese, Romanian, Sicilian, Venetian, Walloon; **West Slavic**: Czech, Polish, Slovak; **Armenian**: Armenian; **Italic**: Latin; **North Germanic**: Danish, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmal), Swedish; **Baltic**: Lithuanian, Latvian; **Indo-Aryan**: Fiji Hindi, Marathi, Urdu, Bosnian, Croatian, Punjabi, Serbian; **East Slavic**: Belarusian, Russian, Ukrainian; **Frisian**: West Frisian

Table 3.1: Table of Indo-European languages used, language families in bold.

Austronesian: Minang, Amharic, Indonesian, Malay, Sundanese, Cebuano, Tagalog, Waray-Waray, Buginese, Javanese; **Altaic**: Mongolian, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Turkish, Tatar, Uzbek; **creole**: Haitian; **Austroasiatic**: Vietnamese; **Kartvelian**: Georgian; **Niger-Congo**: Swahili, Yoruba; **Vasonic**: Basque; **Afro-Asiatic**: Malagasy; **Quechuan**: Quechua; **Semitic**: Arabic, Egyptian Arabic, Hebrew; **Korean**: Korean; **Uralic**: Estonian, Finnish, Hungarian; **Tai**: Thai;

Table 3.2: Table of non-Indo-European languages used, language families in bold.

3.2 Method

We used the lexicons of 97 languages extracted from Wikipedia. The details on these lexicons, including the typological details and our corpus cleaning procedure, are explained in Appendix 3.5. The languages analyzed included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, 12 language families are represented as well as a Creole. The languages analyzed are shown in Tables 3.1 and 3.2.

For this analysis, we defined a lexicon as the set of the 20,000 most frequent unique orthographic wordforms (word types) in a given language. We used only orthographic wordforms here, which are a good proxy for phonological forms (an assumption tested in Section 3.3.2 for a small number of languages).

To assess whether the Wikipedia corpus (which uses orthographic forms and contains morphologically complex words) is a good proxy for a more controlled corpus that uses phonemic representations and is restricted to monomorphemic words, we also analyzed phonemic lexicons derived from CELEX for Dutch, English and German (R. Baayen et al., 1995) and Lexique for French (New, Pallier, Brysbaert, & Ferrand, 2004). The lexicons were restricted to include only monomorphemic lemmas (coded as "M" in CELEX; I.D. (a French native speaker) identified mono-morphemes by hand for French). That is, they contained neither inflectional affixes (like plural -*s*) nor derivational affixes like -*ness*. In order to focus on the most used parts of the lexicon, we selected only words whose frequency is greater than 0. (The CELEX database includes some rare words listed as having 0 frequency, which were not in the original CELEX sample.) Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., 'bat') we included this form only once.

All three CELEX dictionaries were transformed to make diphthongs into 2-character strings. In each lexicon, we removed a small set of words containing foreign characters. This resulted in a lexicon of 5459 words for Dutch, 6512 words for English, 4219 words for German and 6782 words for French.

Variables under consideration:

For each word in each language we computed the word's:

- Length in characters (so that we can compare only words of the same length)
- Token frequency: for orthographic lexicons: across all the Wikipedia corpus of the language; for phonemic lexicons: using the frequency in CELEX
- Orthographic probability (as a proxy for phonotactic probability): We trained an ngram model on characters ($n = 3$ with a Laplace smoothing of 0.01 and with Katz backoff in order to account for unseen but possible sound sequences) on each lexicon and used the resulting model to find the probability of each word string under the model. Table 3.3 shows examples of high and low probability English words under the English language model.

word	log probability
shed	-3.75
reed	-3.69
mention	- 4.63
comment	-4.68
tsar	-8.64
Iowa	-9.47
tsunami	- 12.90
kremlin	-11.53

Table 3.3: Phonotactically likely and unlikely words in English with their log probabilities

- Orthographic neighborhood density (as a proxy for phonological neighborhood density): PND is defined for each word as the number of other words in the lexicon that are one edit (an insertion, deletion, or substitution) away in phonological space (Luce, 1986; Luce & Pisoni, 1998). For instance, 'cat' and 'bat' are phonological neighbors, as well as minimal pairs since they have the same number of letters and differ by 1. 'Cat' and 'cast' are neighbors but not minimal pairs. We will focus on minimal pairs, as opposed to neighbors, in order to avoid confounds from languages having different distributions of word lengths.

3.3 Results

3.3.1 Large-scale effects of frequency on 97 languages

Correlational analysis

Figure 3-1 shows correlations for each language and length (from 4 to 6 letters) separately, between (a) orthographic probability and frequency and (b) minimal pairs and frequency for non-Indo-European languages. Points to the right of the dotted line at 0 show a positive correlation. Almost all languages indeed show a positive correlation. Figure 3-2 shows a similar plot for Indo-European languages. Once again, almost all languages show a positive correlation.

Analyzing each length separately and focusing on words of 3 to 7 letters, we found

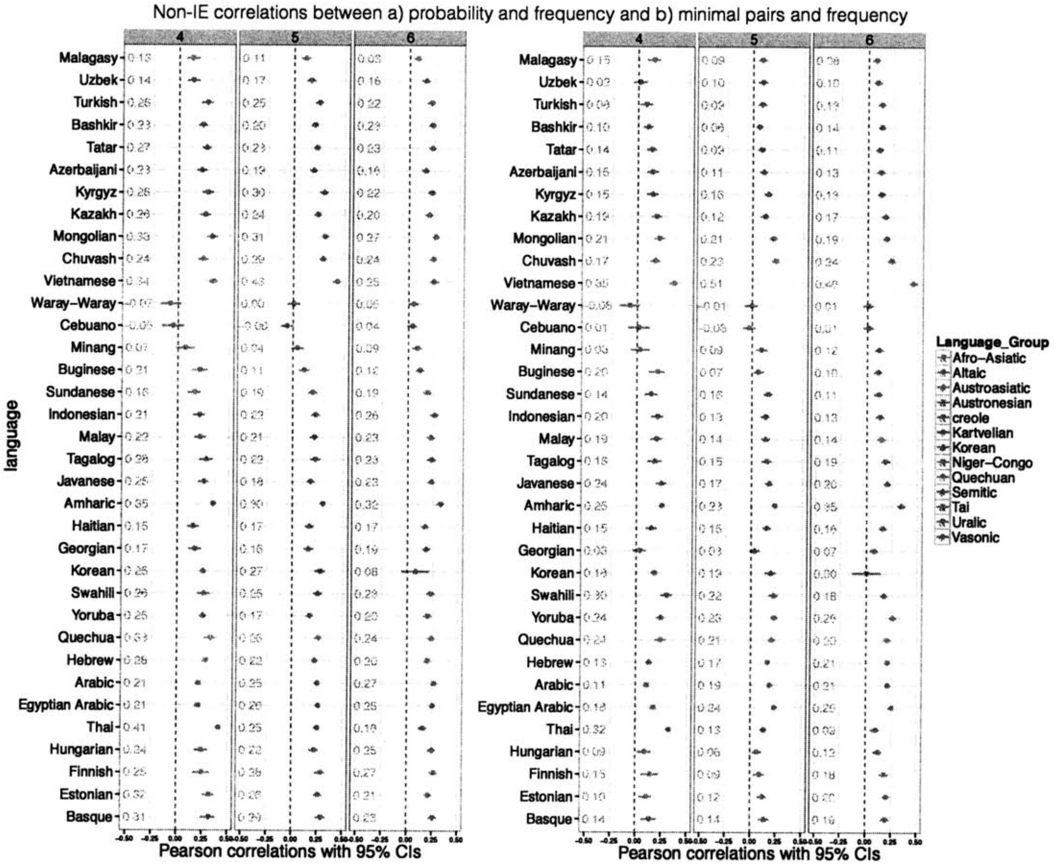


Figure 3-1: Correlation coefficients between a) orthographic probability and frequency and b) minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.

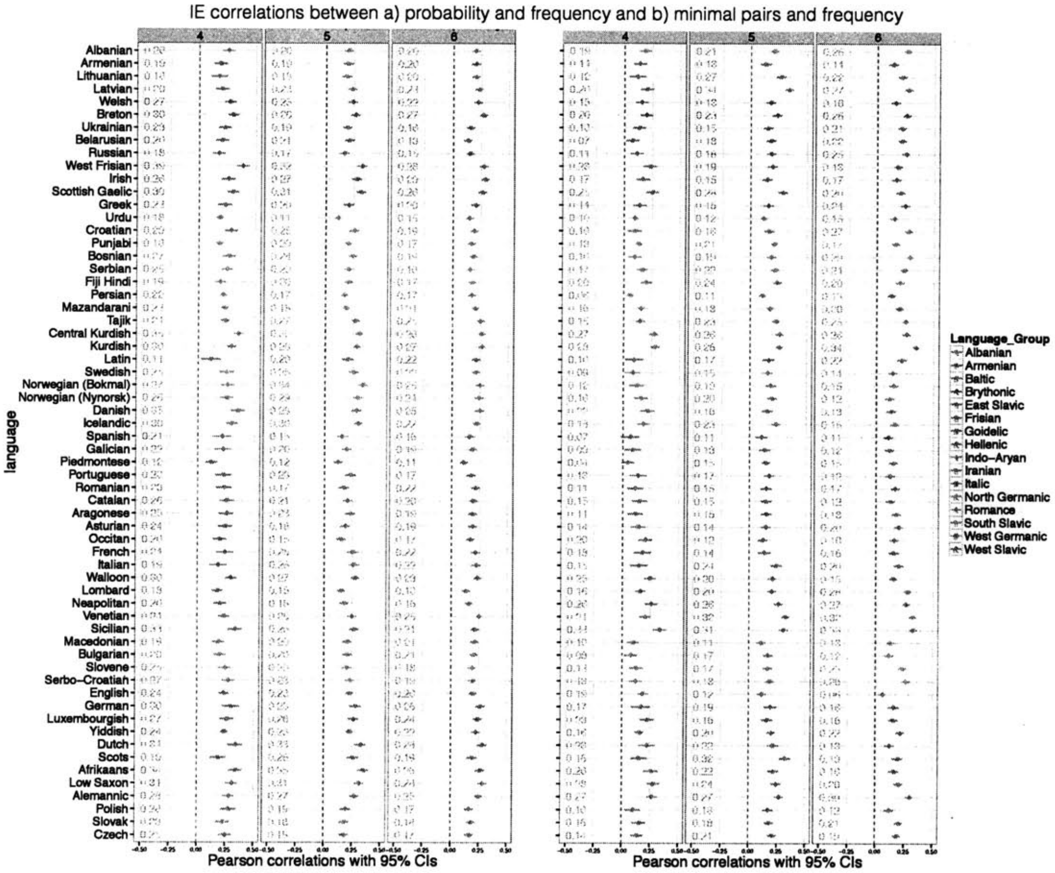


Figure 3-2: Correlation coefficients between a) orthographic probability and frequency and b) minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.

word length	mean correlation	proportion showing positive correlation	proportion showing significant correlation
3 letters	.27 (.26)	1.00 (1.00)	.97 (.92)
4 letters	.24 (.24)	.98 (.98)	.97 (.96)
5 letters	.23 (.22)	.99 (.99)	.98 (.96)
6 letters	.21 (.20)	1.00 (1.00)	.97 (.98)
7 letters	.19 (.19)	1.00 (1.00)	.98 (.99)

Table 3.4: Summary of relationship between orthographic probability and frequency, across languages. Separated by length, (a) the mean correlation across languages for the relationship between orthographic probablility and frequency, (b) the proportion of languages that show a positive correlation between orthographic probability and frequency, and (c) the proportion of languages for which this relationship is significantly different from 0 at $p < .05$. In parentheses, we include each value for the subset of the lexicons that do not appear in the English Subtlex subtitles corpus.

word length	mean correlation	proportion showing positive correlation	proportion showing significant correlation
3 letters	.19 (.19)	1.00 (.99)	.97 (.84)
4 letters	.17 (.16)	.98 (.99)	.97 (.93)
5 letters	.18 (.18)	.98 (.98)	.98 (.96)
6 letters	.19 (.18)	1.00 (1.00)	.97 (.97)
7 letters	.18 (.18)	.99 (.99)	.98 (.96)

Table 3.5: Summary of relationship between minimal pairs and frequency, across languages. Separated by length, (a) the mean correlation across languages for the relationship between number of minimal pairs and frequency, (b) the proportion of languages that show a positive correlation between number of minimal pairs and frequency, and (c) the proportion of languages for which this relationship is significantly different from 0 at $p < .05$. In parentheses, we include each value for the subset of the lexicons that do not appear in the English Subtlex subtitles corpus.

a significant correlation between log frequency and orthographic probability in most languages (see Table 3.4). For instance for the 4-letter words, 96 of 97 languages showed a positive correlation and 95 out of the 97 correlations were significantly positive at $p < .05$.

We also found a robust correlation between log frequency and number of minimal pairs (mean $r = .18$, averaging across the separate correlations found by length and languages) for almost all languages, as shown in Table 3.5.

In order to ensure that any observed effects are not the product of English overlap, we ran the same analyses on the full lexicons as well as on subsets of lexicons that exclude any word that also appears in the English Subtlex subtitles database (Ferrer-i Cancho & Moscoso del Prado Martín, 2011). This excludes English intrusions but also excludes perfectly good words like *die* in German (which means "the" and is unrelated to English "die") and French *dire* (meaning "to say" and unrelated to the English adjective *dire*). Note that, for all lengths, the results obtained when excluding all English words are similar in terms of overall correlation, as can be seen in Table 3.5. Because most of the English words excluded are actually not intrusions but are native words that just happen to also be English forms, we include them in all subsequent analyses. Note that this method does not account for the possibility of borrowings in other languages (and indeed many less widely spoken languages will borrow words from nearby major languages and these borrowings may have different phonotactics). We consider this phenomenon, however, to be part of the natural evolution of language and do not attempt to exclude it. In excluding English, we primarily seek to exclude words that appear in Wikipedia due to noise from HTML intrusions, bad translation, etc.

Additionally, we find a robust correlation between orthographic probability and number of minimal pairs (mean $r = .49$ when we average across the correlations found for each length). This result holds for all lengths across the vast majority of languages and is consistent with the idea that words with high orthographic probability are more likely to have neighbors since the orthographic probabilities of their neighbors will be on average high too. For example, a word like 'set' is more likely to have more minimal pairs in English than the word 'quiz' simply because the letters in 'set' are more common and so, probabilistically, there are more opportunities for a word to be orthographically close to 'set' than to 'quiz.'

It follows that the correlations between frequency and phonological similarity uncovered previously should be (partly) due to both frequency and orthographic probability being correlated with phonological similarity. Thus, the question becomes (a) whether the correlation between frequency and phonological similarity remains after factoring out the effect of orthographic probability and (b) whether the correlation between frequency and orthographic probability remains after factoring out the effect of phonological similarity. Moreover, many of the languages in this study are highly related, so we need an analysis that generalizes across families and languages to make sure that the effect is not just lineage-specific.

Mixed effect analysis

We ran a mixed effect regression predicting (scaled) frequency for each word from orthographic probability and number of minimal pairs, where both predictors were normalized for each language and length. We used a maximal random effect structure with random intercepts for each language, language sub-family, and language family and slopes for orthographic probability and number of minimal pairs for each of those grouping factors. In effect, this random effect structure allows for the possibility that some languages or language families show the predicted effect whereas others do not. It allows us to test whether the effect exists beyond just language-specific trends. Because of the complex random effect structure and the large number of data points, we fit each length separately and focused on words of length 3 through length 7.

For 4-letter words (a representative length), a 1 standard deviation increase in orthographic probability was predictive of a .20 standard deviation increase in frequency; a 1 standard deviation increase in number of minimal pairs was predictive of a .06 standard deviation increase in frequency. To assess the significance of orthographic probability above and beyond the number of minimal pairs, we performed a likelihood ratio test comparing the full model to an identical model without a fixed effect for orthographic probability (but the same random effect structure). The full model was significantly better by a chi-squared test for goodness of fit ($\chi^2(1) = 30.9$, $p < .0001$). To assess the significance of the number of minimal pairs above and beyond the effect of orthographic probability, we compared the full model to an identical model without a fixed effect for the number of minimal pairs using a likelihood ratio test. Once again, the full model explained the data significantly better ($\chi^2(1) = 10.6$, $p < .001$). Thus, both the number of minimal pairs and orthographic probability appear to make independent contributions in explaining word frequency. This effect holds above and

word length	orthographic probability	number of mininimal pairs
3 letters	.23**	.08**
4 letters	.20***	.06***
5 letters	.19***	.07**
6 letters	.15***	.11***
7 letters	.13***	.11***

Table 3.6: Separated by length, the model coefficient from the full model including random intercepts and slopes for language, sub-family, and family for orthographic probability and number of minimal pairs. Two asterisks means that by a likelihood test, the predictor significantly improves model fit at $p < .01$. Three asterisks means $p < .001$.

beyond effects of language family or sub-family, which are included in the model as random effects. Note that the effect size is larger for orthographic probability than it is for number of minimal pairs and that a model including a fixed effect of probability but not minimal pairs has a better model fit (AIC = 520310) than one that includes minimal pairs but not probability as a fixed effect (AIC = 520330). We find a similar pattern of results for all other lengths examined, as summarized in Table 3.6. Overall, these results suggest that both the number of minimal pairs and the orthographic probability independently predict frequency but that the effect of orthographic probability is stronger and is likely, in part, driving the neighborhood effect.

In Appendix 3.6, we show the results of a lasso regression (Tibshirani, 1996), for each length and language, predicting scaled frequency from scaled orthographic probability and scaled number of minimal pairs. As with our other analyses, this analysis suggests that more frequent words have higher orthographic probability and more minimal pairs but that the minimal pairs result is driven, at least in part, by orthographic probability.

3.3.2 Testing correlation generalizability to phonemic representations

We used orthographic lexicons because they could be easily extracted for a large number of languages. However, a better measure of phonotactics could be calculated on the phonemic transcription of words, and a better measure of phonological similarity should exclude morphological similarity by focusing only on monomorphemes. To assess whether the correlation between frequency and phonological similarity and between frequency and phonotactic probability hold in a set of monomorphemic words with phonemic representations, we performed the same analysis using the four phonemic lexicons from Dutch, English, French, and German.

As before, we tested whether the token frequency could be predicted by phonotactic probability (here approximated by *phonemic* probability using a ngram model operating over triphones) and/or number of minimal pairs. The correlations obtained in these four phonemic lexicons replicated previous correlations with the orthographic lexicons for these languages: all four languages still showed positive correlations for the relationship between phonotactic probability and frequency and between number of minimal pairs and frequency.

That said, the correlations were slightly lower in the more controlled set for the 4 languages than when using the same measures in the larger data set: the correlation between minimal pairs and frequencies (across the 4 languages and word lengths 3-7) is, on average, .03 lower for the correlation between minimal pairs and frequency and .10 lower for the correlation between orthographic probability and frequency. This suggests that part of the effect could be driven by morphology–which is absent in the controlled phonemic lexicons but present in the Wikipedia corpus.

3.4 Discussion

We found that frequent wordforms are more likely to be similar to other wordforms and composed of more likely sequence of phonemes than infrequent ones. These correlations were robustly present across a large number and wide variety of typologically different languages. Just as the Zipfian word frequency distribution allows for functional optimization of word lengths (Piantadosi et al., 2011; Piantadosi, 2014) and word forms (Piantadosi, Tily, & Gibson, 2012), this work shows that the frequency profile of even words of the same length is structured in a non-arbitrary way so as to maximize the use of "good" word forms.

Note that, form a strict noisy channel perspective, there is a tradeoff in structuring the lexicon this way. In a language where orthographically probable strings in dense neighborhoods are the most frequent words, there may be an increased chance of perceptual confusion. We found no evidence that lexicons are optimized in this way. On the other hand, there may well be a communicative advantage to having the most frequent words live in dense neighborhoods since those are precisely the words that are easiest to recover from context. So any potential noise could be corrected contextually. We did indeed find a positive correlation between phonological neighborhood density *independent* of our measure of phonotactic probability–lending possible support to this idea. However, it is also possible that the correlation reflects phonotactic effects not capture by our simple orthographic model. More work is needed to assess this possibility.

We do not believe that the main result of this paper is purely a result of morphological regularity since the same analyses run on monomorphemic words in a subset of languages show the same pattern of results. Moreover, although phonotactic constraints are an obvious and major source of regularity in the lexicon, it is important to note that these results are not likely just the result of phonotactic constraints since the results hold even after controlling for the influence of phonotactic probability, at least in the analyses reflecting the influence of word usage. In a companion study (Mahowald et al., *submitted*) in which we constructed a phonotactically-controlled baseline for lexicons, we provided compelling evidence that natural lexicons are more tightly clustered in phonetic space than would be expected by chance, over and above the constraints imposed by phonotactics. This, taken together with the present results, suggests that languages tend to favor wordform similarity in the lexicon.

In this study, we addressed the issue of why and how the structures of lexicons diverges from what can be expected by chance alone, but we leave it to future work to investigate how it got to be that way. One promising body of work in that vein concerns language evolution. Indeed, there has been much experimental work studying the evolution of language showing that language users will preferentially discard forms and structures that are disadvantageous in favor of other, fitter words and phrases (K. Smith, Kirby, & Brighton, 2003; Fedzechkina et al., 2012). Thus, one plausible mechanism for the effects described here is that generations of learners improve on the lexicon, honing it over time by avoiding words that are too strange, complex, or that otherwise don't fit with the rest of the words in the lexicon.

In summary, across 97 typologically different languages we found that frequent words are likely to be more orthographically similar to one another and composed of more common letter sequences than infrequent words. These effects suggests that the structure of wordforms in the lexicon is functionally motivated across a wide range of typologically diverse languages and therefore that functional considerations are a major factor in the architecture of human language.

3.5 Appendix: Dataset of 97 lexicons from Wikipedia

We started with lexicons of 115 languages from their Wikipedia databases (`https://dumps.wikimedia.org`). We then excluded languages for which a spot-check for non-native (usually English) words in the top 100 most frequent words in the lexicon between 3 and 7 characters revealed more than 80% of words were not native. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the 3-7 letter words in Chinese Wikipedia are often English. However, we included languages like Korean in which words generally consist of several characters. After these exclusions, 97 languages remained.[1] We analyzed the data both with and without these exclusions, and the exclusions do not significantly affect the overall

[1] We excluded: Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, Kannada

direction or magnitude of the results. The final languages included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, there are 12 language families represented as well as a Creole.

To get a sense of how clean these Wikipedia lexicons are, we randomly sampled 10 languages for which we then inspected the 100 most frequent words and an additional 100 random words to look for intrusion of English words, HTML characters, or other undesirable properties.

For the top 100 words in the lexicons of the 10 sampled languages, we found at most 3 erroneous words. For the same languages, we also inspected a randomly selected 100 words and found that the mean number of apparently non-intrusive words was 93.5 (with a range from 85 to 99). The most common intrusion in these languages was English words.

3.6 Appendix: Lasso regression analysis

Lasso analysis

We fit separate lasso regressions (Tibshirani, 1996) for each length and language predicting scaled frequency from scaled orthographic probability and scaled number of minimal pairs. The lasso regression puts a constraint on the sum of the absolute value of the regression coefficients (L1-regularization) and thus effectively pushes some coefficients to 0 if they are not needed. We set the value of the lasso parameter using cross-validation to minimize the out-of-sample error and used the `lars` software package (Efron, Hastie, Johnstone, & Tibshirani, 2004) in R.

In Figure 3-3, for each language, we plot the coefficient for scaled number of minimal pairs on the x-axis against the coefficient for scaled orthographic probability on the y-axis. If both predictors show robust effects, we would predict the points to cluster in the upper right quadrant. If both predictors showed little effect, the points would cluster around 0.

Although the lasso regression drives some of these coefficients to 0, the plot clearly shows effects of both minimal pairs and orthographic probability (with larger coefficients in general for orthographic probability). Specifically, for 4 and 5 letter words, only 2 languages show negative coefficients for orthographic probability and none for 6 and 7 letter words.

Thus, it appears that more frequent words also have higher orthographic probability as well as more minimal pairs, although the minimal pairs effect is driven in part by orthographic probability.

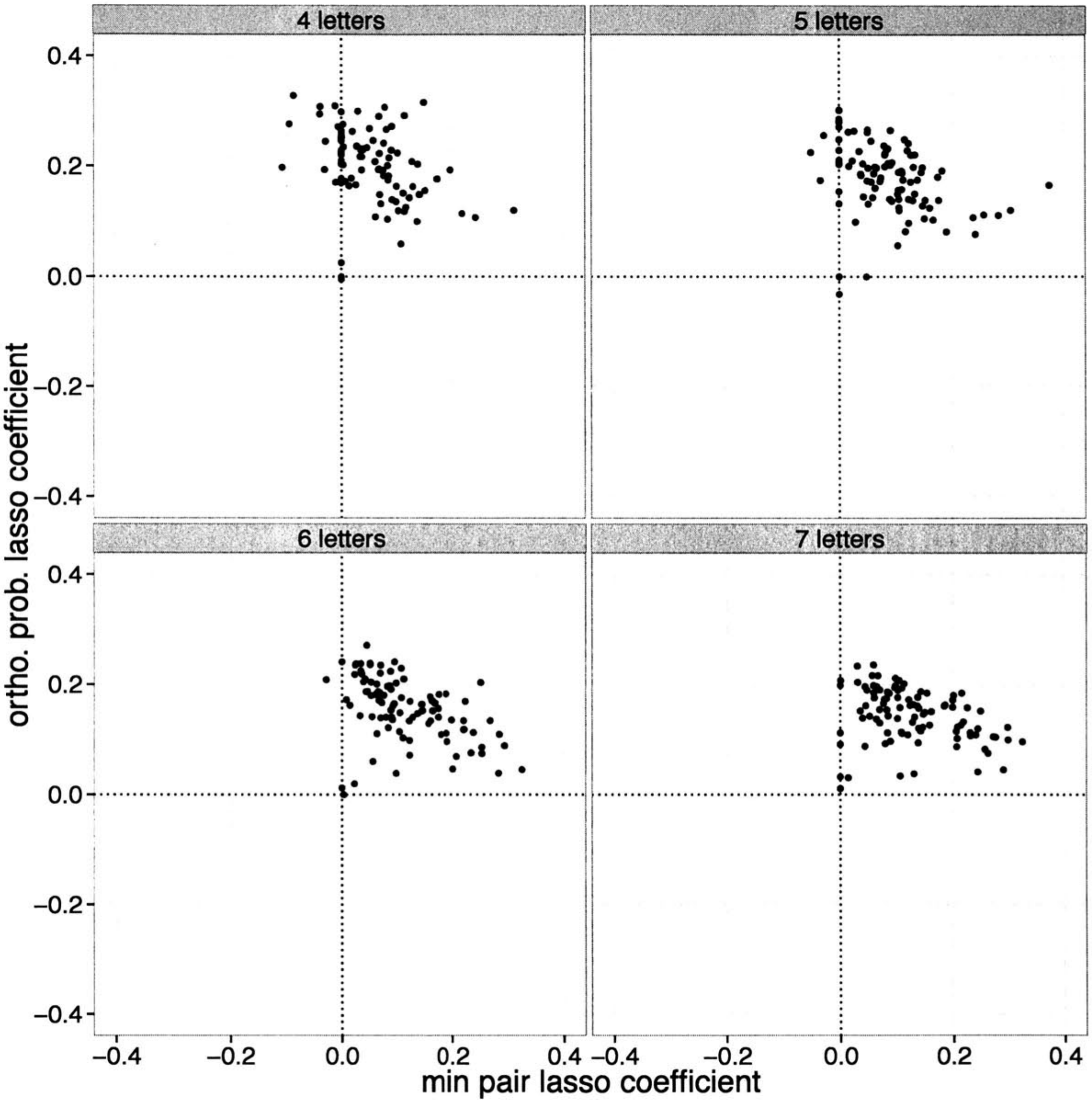


Figure 3-3: This plot shows the lasso regression coefficients (predicting scaled frequency) for scaled number of minimal pairs and scaled orthographic probability.

Chapter 4

Simulating null lexicons

In Chapter 2, we explored how the most frequent words in a language tend to be short. In Chapter 3, we explored how the most frequent words tend to be phonotactically probable and reside in dense neighborhoods. Both of these facts, when pushed to the extremes, suggest that languages are designed to favor word forms that are *similar* to one another. But that raises an obvious question: is it really the case that a well-designed language should have words that are tightly clustered in phonetic space?

Imagine a language that uses the word 'feb' to refer to the concept HOT, and that the language now needs a word for the concept warm. If the language used the word 'fep' for WARM, it would be easy to confuse with 'feb' (HOT) since the two words differ only in the voicing of the final consonant and would often occur in similar contexts (i.e. when talking about temperature). However, the similarity of 'feb' and 'fep' could make it easier for a language learner to learn that those sound sequences are both associated with temperature, and the learner would not have to spend much time learning to articulate new sound sequences since 'feb' and 'fep' share most of their phonological structure. On the other hand, if the language used the word 'zoz' for the concept WARM, it is unlikely to be phonetically confused with 'feb' (HOT), but the learner might have to learn to articulate a new set of sounds and would need to remember two quite different sound sequences that refer to similar concepts.

Here, we investigate how communicative efficiency and learnability trade off in the large-scale structure of natural languages. We have developed a set of statistical tools to characterize the large-scale statistical properties of the lexicons. Our analysis focuses on testing and distinguishing two pressures in natural lexicons: a *pressure for dispersion* (improved discriminability) versus a *pressure for clumpiness* (re-use of sound sequences). Below, we discuss each in more detail.

A pressure for dispersion of wordforms

Under the noisy channel model of communication (Gibson, Bergen, & Piantadosi, 2013; Levy, 2008a; Shannon, 1948), there is always some chance that the linguistic signal will be misperceived as a result of errors in production, errors in comprehension, inherent ambiguity, and other sources of uncertainty for the perceiver. A lexicon is maximally robust to noise when the expected phonetic distance among words is maximized (Flemming, 2004; Graff, 2012), an idea used in coding theory (Shannon, 1948). Such dispersion has been observed in phonetic inventories (Flemming, 2002; Hockett & Voegelin, 1955) in a way that is sensitive to phonetic context (Steriade, 2001, 1997). The length and clarity of speakers' pronunciations are also sensitive to context predictability and frequency (e.g., Bell et al., 2003; Cohen Priva, 2008; Pluymaekers, Ernestus, & Baayen, 2005; Aylett & Turk, 2004; Van Son & Van Santen, 2005; Raymond, Dautricourt, & Hume, 2006), such that potentially confusable words have been claimed to be pronounced more slowly and more carefully. Applying this idea to the set of wordforms in a lexicon, one would expect wordforms to be maximally dissimilar from each other, within the bounds of conciseness and the constraints on what can be easily and efficiently produced by the articulatory system. Indeed, a large number of phonological neighbors (i.e., words that are one edit apart like 'cat' and 'bat') can impede spoken word recognition (Luce, 1986; Luce & Pisoni, 1998), and the presence of lexical competitors can affect reading times (Magnuson et al., 2007). Phonological competition may also be a problem in early stages of word learning: young toddlers fail to use a single-feature phonological distinction to assign a novel meaning to a wordform that sounds similar to a very familiar one (e.g., learning a novel word such as "tog" when having "dog" in their lexicon, (Swingley & Aslin, 2007; Dautriche, Swingley, & Christophe, 2015)).

A pressure for clumpiness of wordforms

Dispersion of wordforms in the lexicon may be functionally advantageous. Yet, it is easy to see that a language with a hard constraint for dispersion of wordforms will have many long, therefore complex, words (as words need to be distinctive). Well designed-lexicon must also be composed of simple signals that are easily memorized, produced, processed and transmitted over generations of learners. In the extreme case, one could imagine a language with only one wordform. Learning the entire lexicon would be as simple as learning to remember and pronounce one word. While this example is absurd, there are several cognitive advantages for processing words that are similar to other words in the mental lexicon. Words that overlap phonologically with familiar words are considered to be easier to process because they receive support from stored phonological representations. There is evidence that words that have many similar sounding words in the lexicon are easier to remember than words that are more phonologically distinct (Vitevitch, Chan, & Roodenrys, 2012) and facilitate production as evidenced by lower speech error rates (Stemberger, 2004; Vitevitch & Sommers, 2003) and naming latencies (Vitevitch & Sommers, 2003) (but see (Sadat, Martin, Costa, & Alario, 2014) for a review of the sometimes conflicting literature on the effect of neighborhood density on lexical production). Additionally, words with many phonological neighbors tend to be phonetically reduced (shortened in duration and produced with more centralized vowels) in conversational speech (Gahl, Yao, & Johnson, 2012; Gahl, 2015).This result is expected if faster lexical retrieval is associated with greater phonetic reduction in conversational speech as it is assumed for highly predictable words and highly frequent words (Bell et al., 2003; Aylett & Turk, 2006). In sum, while words that partially overlap with other words in the lexicon may be difficult to recognize (Luce, 1986; Luce & Pisoni, 1998), they seem to have an advantage for memory and lexical retrieval.

One source of wordform regularity in the lexicon comes from a correspondence between phonology and semantic and/or syntactic factors. Words of the same syntactic category tend to share phonological features, such that nouns sound like nouns, verbs like verbs, and so on (Kelly, 1992). Similarly, phonologically similar words tend to be more semantically similar across a wide variety of languages (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Dautriche, Mahowald, Gibson, & Piantadosi, submitted). The presence of these natural clusters in semantic and syntactic space therefore result in the presence of clusters in phonetic space. Imagine, for instance, that all words having to do with sight or seeing had to rhyme with 'look.' A cluster of '-ook' words would develop, and they would all be neighbors and share semantic meaning. One byproduct of these semantic and syntactic clusters would be an apparent lack of sparsity among wordforms in the large-scale structure of the lexicon. There is evidence that children and adults have a bias towards learning words for which the relationship between their semantics and phonology is not arbitrary (Imai & Kita, 2014; Imai, Kita, Nagumo, & Okada, 2008; Nygaard, Cook, & Namy, 2009; Nielsen & Rendall, 2012; Monaghan, Christiansen, & Fitneva, 2011; Monaghan et al., 2014). However such correspondences between phonology and semantic may affect some aspects of the production system: speech production errors that are semantically and phonologically close to the target (e.g., substituting 'cat' by 'rat') are much more likely to occur than expected than errors than are purely semantic (e.g., substituting 'cat' by 'dog') or purely phonological (e.g., substituting 'cat' by 'mat') in spontaneous speech (Schwartz, Dell, Martin, Gahl, & Sobel, 2006; Goldrick & Rapp, 2002; Dell & Reich, 1981).

Another important source of phonological regularity in the lexicon is *phonotactics*, the complex set of constraints that govern the set of sounds and sound combinations allowed in a language (Hayes & Wilson, 2008; Vitevitch & Luce, 1998). For instance, the word 'blick' is not a word in English but plausibly could be, whereas the word 'bnick' is much less likely due to its implausible onset *bn-* (Chomsky & Halle, 1965).[1] These constraints interact with the human articulatory system: easy-to-pronounce strings like 'ma' and 'ba' are words in many human languages, whereas some strings, such as the last name of Superman's nemesis *Mister Mxyzptlk*, seem unpronounceable in any language.[2]

[1] There are many existing models that attempt to capture these language-specific rules. A simple model is an n-gram model over phones, whereby each sound in a word is conditioned on the previous n-1 sounds in that word. Such models can be extended to capture longer distance dependencies that arise within words (Gafos, 2014) as well as feature-based constraints such as a preference for sonorant consonants to come after less sonorant consonants (Albright, 2009; Goldsmith & Riggle, 2012; Hayes, 2012; Hayes & Wilson, 2008).

[2] Though as a anonymous reviewer pointed out, some have succeeded in doing so (`https://en.wikipedia.org/wiki/Mister_Mxyzptlk#Pronunciation`)

Nevertheless, the phonotactic constraints of a language are often highly language-specific. While English does not allow words to begin with *mb*, Swahili and Fijian do. Phonotactic constraints provide an important source of regularity that aids production, lexical access, memory and learning. For instance, words that are phonotactically probable in a given language (i.e., that make use of frequent transitions between phonemes) are recognized more quickly than less probable sequences (Vitevitch, 1999). Furthermore, infants and young children seem to learn phonotactically probable words before learning less probable words (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer listening to high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013).[3]

The upshot of this regularity for the large-scale structure of the lexicon is to *constrain* the lexical space. For instance, imagine a language called *Clumpish* in which the only allowed syllables were those that consist of a nasal consonant (like *m* or *n*) followed by the vowel *a*. Almost surely, that language would have the words 'ma', 'na', 'mama', 'mana', and so on since there are just not that many possible words to choose from. The lexical space would be highly constrained because most possible sound sequences are forbidden. From a communicative perspective, such a lexicon would be disadvantageous since all the words would sound alike. The result would be very different from the lexicon of a hypothetical language called *Sparsese* in which there were no phonotactic or articulatory constraints at all and in which any phoneme was allowed. In a language like that, lexical neighbors would be few and far between since the word 'ma' would be just as good as 'Mxyzptlk'.

Assessing lexical structure

In this chapter, we ask whether the lexicon is clumpy or sparse. But, because of phonotactics and constraints on the human articulatory system, a naive approach would quickly conclude that the lexicon is clumpy. Natural languages look more like *Clumpish* than they do like *Sparsese* since any given language uses only a small portion

[3] Note that wordform similarity seems to have a different influence on word learning: phonological probability helps learning but neighborhood density makes it difficult to attend to and encode novel words (Storkel, Armbruster, & Hogan, 2006).

of the phonetic space available to human language users.[4] We therefore focus on the question of whether lexicons show evidence for clumpiness or sparsity above and beyond phonotactics in the *overall* (aggregate) structure of the lexicon.

The basic challenge with assessing whether a pressure for dispersion or clumpiness drives the organization of wordform similarity in the lexicon is that it is difficult to know what statistical properties a lexicon should have in their absence. If we believe, for instance, that the wordforms chosen by English are clumpy, we must be able to quantify clumpiness compared to some baseline. Such a baseline would reflect the *null hypothesis* about how language may be structured in the absence of cognitive forces. Indeed, our methods follow the logic of standard statistical hypothesis testing: we create a sample of null lexicons according to a statistical baseline with no pressure for either clumpiness nor dispersion. We then compute a test measure (e.g., string edit distance) and assess whether real lexicons have test measures that are far from what would be expected under the null lexicons. We present a novel method to compare natural lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or scattered wordforms would be as the result of only phonotactics.[5] Across a variety of measures, we find that natural lexicons have the tendency to be clumpier than expected by chance (even when controlling for phonotactics). This reveals a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonetically distinct as possible.

4.1 Method

Assessing the extent to which the lexicons of natural languages are clumpy or sparse requires a model of what wordforms should be expected in a lexicon in the absence of either force. Prior studies looking at the statistics of language—in particular Zipf's

[4] As an illustration, English has 44 phonemes so the number of possible unique 2-phone words is $44^2 = 1936$, yet there is only 225 unique 2-phone word forms in English (among all the word forms appearing in CELEX (R. Baayen, Piepenbrock, & van H, 1993), thus only 11% of the space possible for two-phone words is actually used in English.

[5] Using a similar approach, H. Baayen (1991) studied wordform similarity in relation to words' frequency by simulating lexicons (see also R. Baayen (2001)'s implementation of the Simon-Mandelbrot model.)

law (Mandelbrot, n.d.; Miller, 1957)—have made use of a *random typing* model in which sub-linguistic units are generated at random, occasionally leading to a word boundary when a "space" character is emitted. However, this model makes unrealistic assumptions about the true generative processes of language (Howes, 1968; Piantadosi, Tily, & Gibson, 2013) as the sequences of sounds composing words are not generated randomly but follow complex constraints (Hayes, 2012; H. Baayen, 1991). To more accurately capture the phonotactic processes at play in real language, here we built several generative models of lexicons: ngrams over phones, ngrams over syllables, and a PCFG over syllables. After training, we evaluated each model on a held-out dataset to determine which most accurately captured each language. The best model was used as the statistical baseline with which real lexicons are compared. We studied monomorphemes of Dutch, English, German and French. Because our baseline models capture effects of phonotactics, we are able to assess pressures for clumpiness or dispersion over and above phonotactic and morphological regularities.

4.1.1 Real Lexicons

We used the lexicons of languages for which we could obtain reliably marked morphological parses (i.e., whether a word is morphologically simple like 'glad' or complex like 'dis-interest-ed-ness'). For Dutch, English and German we used CELEX pronunciations (R. Baayen et al., 1993) and restricted the lexicon to all lemmas which CELEX tags as monomorphemic. The monomorphemic words in CELEX were compiled by linguistic students and include all words that were judged to be nondecomposed. For French, we used Lexique (New et al., 2004), and a native French speaker identified monomorphemic words by hand. Note that, for Dutch, French and German, these monomorphemic lemmas include infinitival verb endings (*-er* in French, *-en* or *-n* in German and Dutch).[6] Since it is not clear how to separate homophones from polysemy, we chose to focus on surface phonemic forms: when two words with different

[6] Removing these verb endings and running the same analysis on the roots did not change the results observed for these 3 languages (but see section 4.3.2 for an analysis where verb endings matter)

spellings shared the same phonemic wordform (e.g., English 'pair' and 'pear' are both pronounced /per/), we included that phonemic form only once. This accounted for 236 words in Dutch, 646 words in English and 193 words in German. Note that by discarding these words, we already exclude a source of clumpiness in the lexicon.

In order to focus on the most used parts of the lexicon and not on words that are not actually ever used by speakers, we used only those words that were assigned non-zero frequency in CELEX or Lexique, including these words in the simulation however, does not change the results observed. All three CELEX dictionaries were transformed to turn diphthongs into 2-character strings in order to capture internal similarity among diphthongs and their component vowels. In each lexicon, we removed a small set of words containing foreign characters and removed stress marks. Note that since we removed all the stress marks in the lexicons, noun-verb pairs that differ in the position of stress were counted as a single wordform in our lexicon (e.g., in English the wordform 'desert' is a noun when the stress in on the first vowel 'désert' but is a verb when the stress is on the last vowel 'desért' but we use only the wordform /desert/ once). This resulted in a lexicon of 5343 words for Dutch, 6196 words for English, 4121 words for German and 6728 words for French.

4.1.2 Generative models of Lexicons

In order to evaluate each real lexicon against a plausible baseline, we defined a number of lexical models. These models are all generative and commonly used in natural language processing (NLP) applications in computer science. The advantage of using generative models is that we can use the set of words of real lexicons to construct a probability distribution over some predefined segments (phones, syllables, etc.) that can be then used to generate words, thus capturing phonotactic regularities.[7] These models are all lexical models, that is, their probability distributions are calculated using word types as opposed to word tokens, so that the phonemes or the syllables

[7] Fine-grained models of phonotactics exist for English (e.g., Hayes (2012)) yet adapting them to other languages is not straightforward and there is no common measure that will allow us to compare their performances.

from a frequent word like *the* are not weighted any more strongly than those from a less frequent word.[8] We defined three categories of models:

- **n-phone models**: For n from 1 to 6, we trained a language model over n phones. Like an n-gram model over words, the n-phone model lets us calculate the probability of generating a given phoneme after having just seen the previous *n-1* phonemes: $P(x_i|x_{i-(n-1)},...,x_{i-1})$. The word probability is thus defined as the product of the transitional probabilities between the phonemes composing the word, including symbols for the beginning and end of a word. For example, the word 'guitar' is represented as ▶ g ɪ t ɑː r ◀ in the lexicon where ▶ and ◀ are the start and the end symbols. The probability of *guitar* considering a bigram model is therefore:

  $$P(\text{g}|\blacktriangleright) \times P(\text{ɪ}|\text{g}) \times P(\text{t}|\text{ɪ}) \times P(\text{ɑː}|\text{t}) \times P(\text{r}|\text{ɑː}) \times P(\blacktriangleleft|\text{r})$$

  These probabilities are estimated from the lexicon directly. For example $P(\text{ɑː}|\text{t})$ is the frequency of tɑː divided by the frequency of t.

- **n-syll models**: For n from 1 to 2, we trained a language model over syllables. Taking the same example as above, 'guitar' is represented as ▶ gɪ tɑːr ◀ and its probability from a bigram novel over syllable is:

  $$P(\text{gɪ}|\blacktriangleright) \times P(\text{tɑːr}|\text{gɪ}) \times P(\blacktriangleleft|\text{tɑːr})$$

  In order to account for out-of-vocabulary syllables in the final log probabilities, we gave them the same probability as the syllables appearing one time in the training set.

- **Probabilistic Context Free Grammar** (PCFG; Manning and Schütze (1999)): Words are represented by a set of rules of the form $X \rightarrow \alpha$ where X is a non-

[8] Admittedly, the experience speakers have of real language is token-based, and not type-based. Yet, using token-based probability estimates instead of type-based probability estimates to capture phonotactic regularities does not change the pattern of results for the 4 languages.

terminal symbol (e.g., Word, Syllable, Coda) and α is a sequence of symbols (non-terminal and phones). We defined a word as composed of syllables differentiated by whether they are initial, medial, final or both initial and final.

$$\begin{aligned}
Word &\rightarrow SyllableI\ (Syllable)^{+}\ SyllableF \\
Word &\rightarrow SyllableIF \\
Syllable &\rightarrow (Onset)\ Rhyme \\
Rhyme &\rightarrow Nucleus\ (Coda) \\
Onset &\rightarrow Consonant^{+} \\
Nucleus &\rightarrow Vowel^{+} \\
Coda &\rightarrow Consonant^{+}
\end{aligned}$$

These rules define the possible structures for words in the real lexicon. They are sufficiently general to be adapted to the four languages we are studying, given the set of phonemes for each language. Each rule has a probability that determines the likelihood of a given word. The probabilities are constrained such that for every non-terminal symbol X, the probabilities of all rules with X on the left-hand side sum to 1: $\sum P(X \rightarrow \alpha) = 1$. The likelihood of a given word is thus the product of the probability of each rule used in its derivation. For example, the likelihood of 'guitar' is calculated as the product of all probabilities used in the derivation of the best parse (consonant and vowel structures are not shown for simplification):

```
Word → SyllableI(Onset(g) Rhyme(Nucleus(ɪ)))
SyllableF(Onset(t) Rhyme(Nucleus(ɑː) Coda(r)))
```

The probabilities for the rules are inferred from the real lexicon using the Gibbs sampler used in Johnson, Griffiths, and Goldwater (2007) and the parse trees for each word of the held-out set are recovered using the CYK algorithm (Younger, 1967).

4.1.3 Selection of the best model

To evaluate the ability of each model to capture the structure of the real lexicon, we trained each model on 75% of the lexicon (the training set) and evaluated the probability of generating the remaining 25% of the lexicon (the validation set). This process was repeated over 30 random splits of the dataset into training and validation sets. For each model type, we smoothed the probability distribution by assigning non-zero probability to unseen ngrams or rules in the case of the PCFG. This was to allow us to derive a likelihood for unseen but possible sequences of phonemes in the held-out set. Various smoothing techniques exist, but we focus on Witten-Bell smoothing and Laplace smoothing which are straightforward to implement in our case.[9] All smoothing techniques were combined with a backoff procedure (though not for the PCFG), such that if the context AB of a unit U has never been observed ($p(U|AB) = 0$) then we can use the distribution of the lower context ($p(U|B)$). The smoothing parameter was set by doing a sweep over possible parameters and choosing the one that maximized the probability of the held-out set. The optimal smoothing was obtained with Laplace smoothing with parameter .01 and was used in all models described.

In order to compare models, we summed the log probability over all words in the held-out set. The model that gives the highest log probability on the held-out data set is the best model, in that it provides a "best guess" for generating random lexicons that respect the phonotactics of the language.

As shown in Figure 4-1, the 5-phone model gives the best result for all lexicons. In all cases, the 6-phone was the next best model, and the 4-phone was close behind, implying that n-phone models in general provide an accurate model of words. The syllable-based models performed particularly poorly. Thus, we focus our attention on the 5-phone model in the remainder of the results, treating this as our best guess about the null structure of the lexicon.

[9] Other smoothing techniques such as Good Turing or Kneser-Ney cannot be implemented easily as they rely on the number of units for which frequency is equal to one which is not available in every model we tested.

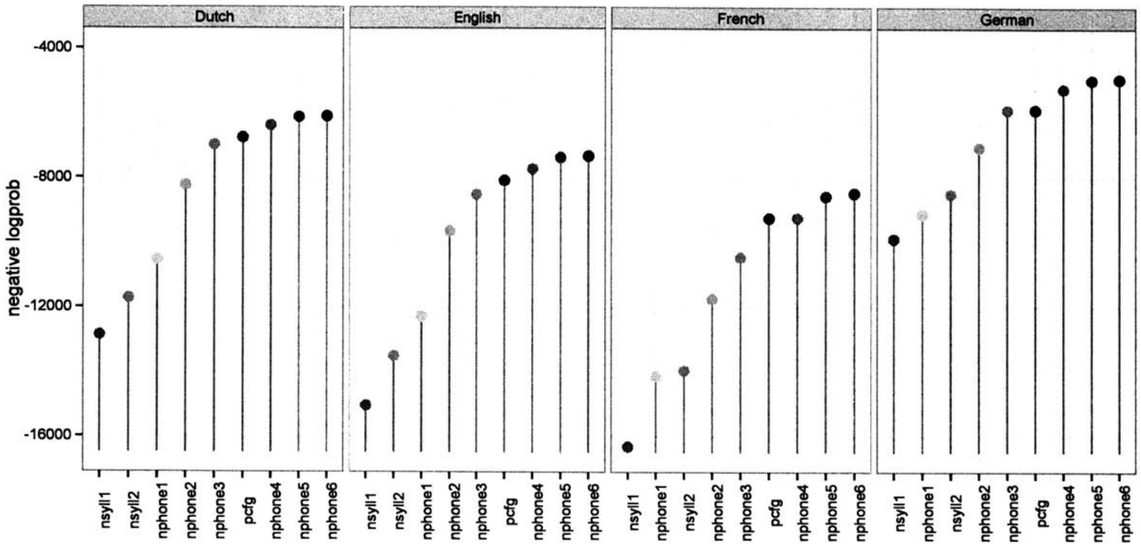


Figure 4-1: Each point represent the mean log probability of one model to predict the held-out data set. The nphone models are represented in green, the nsyll models in pink and the PCFG in blue. The 5-phone model has the highest log probability (indicated by a red segment) for all languages. Standard deviation of the mean are represented but too small to be visible at this scale.

4.1.4 Building a baseline with no pressure for clumpiness or dispersion

We use the 5-phone model to generate simulated null lexicons—ones without any pressure for clumpiness or dispersion other than the 5-phone generating process—and study the position of the real lexicon with respect to the simulated ones. For each language, we trained the 5-phone model on the entire real lexicon and used the resulting language model to generate words for 30 simulated lexicons. It is simplest to visualize how word generation works for the 1-phone case. In such case, all the phones of a given language cover the entire probability space from 0 to 1, each phone covering an interval proportional to its frequency in the real lexicon. We pick a random number between 0 and 1 and and select the phone that corresponds to that value. Phones are generated until the we randomly generate the end-symbol. For the 5-phone model, the same technique is applied except that each phone generation is constrained by the last 4-phones of the word: We first generate a random 5-phone sequence starting with 4 start-symbols, then we generate the next 5-phone sequence to follow given the last 4 phones of the word according to the sequence probability, and so on until the end-symbol is met.

The number of words generated for each simulated lexicon matched the number of words in the corresponding real lexicon. We additionally constrained the generation to ensure that the distribution of word lengths in each simulated lexicon matches the distribution of word lengths in the real lexicon and that, similarly to real lexicons, the simulated lexicon contained no homophones. Practically, it means that we discarded a word every time we generated a word that was not matching the distribution of word lengths of the real lexicon (either because all words of that length have already been generated, or that word's length does not exist in the real lexicon) or was already existing in the simulated lexicon.

On average our best lexicon model generated 52% real words for Dutch, 53% for English, 47% for French, and 41% for German. Note that it is not surprising that the best model generates *only* about 50% real words since the smoothing parameter allowed the generation of non-words likely to be attested in the language.

4.2 Results: Overall similarity in the lexicon

To compare real and simulated lexicons, it is necessary to define a number of test statistics that can be computed on each lexicon to assess how it uses its phonetic space. As in null hypothesis testing, we compute a z-score using the mean and standard deviation estimated from 30 lexicons generated by our best lexicon model. We then ask whether the real lexicon value falls outside the range of values that could be expected by chance under the null model. The p-value reflects the probability that the real lexicon value could have arisen by chance under our chosen 5-phone null model.

We present result separately for a number of different measures of wordform similarity.

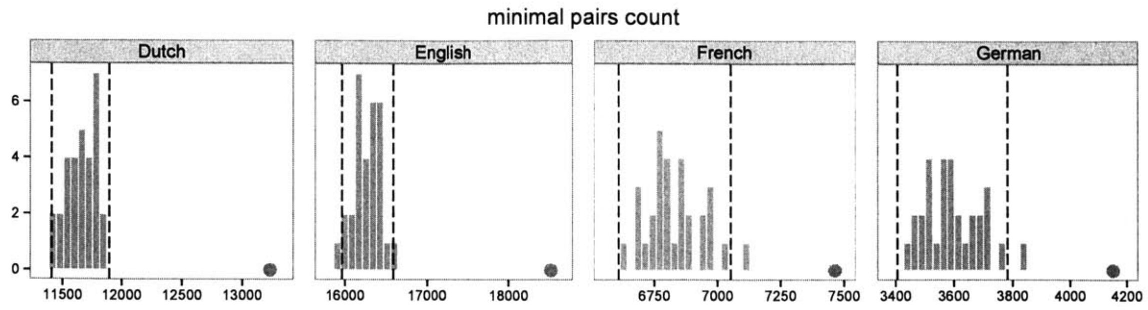


Figure 4-2: Comparison of the total number of minimal pairs for each language (red dot) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. For all four languages, the real lexicon has significantly more minimal pairs than predicted by our baseline.

4.2.1 Minimal pairs

We first considered the number of minimal pairs present in each lexicon. A minimal pair is a pair of words of the same length for which a single sound differs (e.g., 'cat' and 'rat'). If real lexicons are clumpier than expected by chance, then the real lexicons should have more minimal pairs than their simulated counterparts. If they are more dispersed, the real lexicons will have fewer minimal pairs.

Figure 4-2 summarizes this hypothesis test, showing how the various simulated lexicons compare to the real lexicons in terms of number of minimal pairs for each language. Each histogram represents a distribution of minimal pair counts broken up by language across the 30 simulated lexicons. The red dot represents the real lexicon value and the dotted lines represent the 95% confidence interval. All histograms fall to the left of the red dot, which suggests that the real lexicon has more minimal pairs than any of the simulated ones in all four languages (all *ps* < .001; see Table 4.1). This pattern suggests that the real lexicon is clumpier than expected by chance.

To see whether this effect is driven by words of specific length, we looked at the number of minimal pairs for each length. We concentrated on words of length 2 to 7 which represent more than 90% of all words in each language. As shown in Figure 4-3, the real lexicon has more minimal pairs than the simulated ones consistently across words of any length. For all languages, the effect is larger for words of smaller

	Dutch	English	French	German
real	13,237	18,508	7,464	4,151
μ (simulated)	11,653	16,276	6,830	3,594
σ (simulated)	124	159	113	96
z	12.77	14.03	5.61	5.80
p	**<.001**	**<.001**	**<.001**	**<.001**

Table 4.1: z-statistics comparing the total number of minimal pairs in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of minimal pairs counts in the 30 simulated lexicon for each language.

length (length 3 to 4; 30 to 50% of all words in each language) where most minimal pairs are observed. The smaller effect for longer words (especially words of length 7 and above) is likely due to a floor effect since longer words are far less likely to have minimal pairs than short words. Note that, for words of length 2, we see a somewhat degenerate case since there are relatively few possible 2-phoneme words, yet for at least 3 languages it appears that there are more minimal pairs of length 2 than what would be expected by chance. This is explained by the smoothing parameter of the model that allows the generation of unseen sequences of sounds (recall that we smoothed the probability distribution to account for rare sequences of sound that may be unseen in the lexicon of monomorphemes). As a result the model is not exactly reproducing all the 2-phoneme words of the languages.[10]

4.2.2 Levenshtein distance

We can evaluate clustering using more global measures by considering the average string edit distance (*Levenshtein distance*) between words (Levenshtein, 1966). The Levenshtein distance between two sound strings is simply the number of insertions, deletions and replacements required to get from one string to another. For instance, the Levenshtein distance between 'cat' and 'cast' is 1 (insert an 's'), and it is 2 between 'cat' and 'bag' (c → b, t →g). To derive a measure of Levenshtein distance

[10] Inspection of these 2-phoneme words reveals that most of these words are actual wordforms present in the language (hence attested forms, e.g. *is* in English) but are not counted as distinct monomorphemic lemmas and thus are not included in our real lexicons.

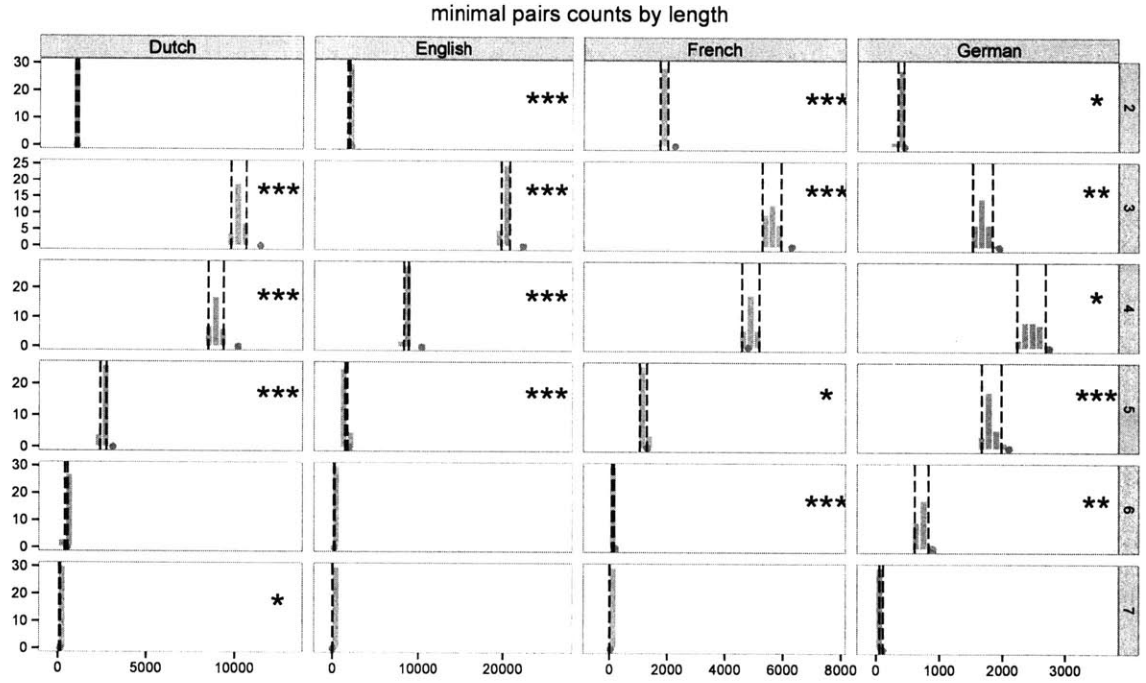


Figure 4-3: Comparison of the number of minimal pairs by word length (2-7) for each language (red dots) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. One star represents $p < .05$, two stars $p < .01$, and three stars $p < .001$.

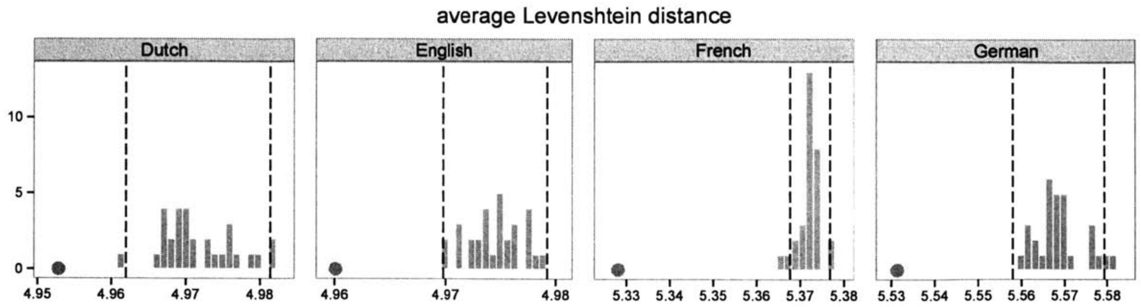


Figure 4-4: Distribution of average Levenshtein distances for each of the 30 simulated lexicons. The red dot represents the real lexicon's value, and the dotted lines are 95% confidence intervals.

that summarizes the whole lexicon, we compute the *average Levenshtein distance* between words in the lexicon by simply computing the distance between every pair of words in the lexicon and then averaging these distances.[11] If the lexicon is clumpier than expected by chance, words will tend to be more similar to one another and we expect to observe a smaller average Levenshtein distance. In contrast, a larger average Levenshtein distance in the real lexicons relative to the simulated lexicons would suggest that the lexicon is more dispersed than expected by chance.

As shown in Figure 4-4, the average Levenshtein distance between words is significantly smaller for the real lexicon than in the simulated lexicons for all four languages (see Table 4.2). The difference is numerically small, but that is to be expected because minimal pairs are statistically unlikely. That is, the edit distance between two words is largely a product of their lengths. For example, on average, the edit distance between two 5-letter words is 5. Nonetheless, the Levenshtein metric provides us with an additional piece of evidence that words in the real lexicons are more similar to each other than what would be expected by chance.

Similarly, to see whether this effect is driven by words of specific length, we looked at the average Levenshtein distance for words of length 2 to 7. As shown in Figure

[11] A possible objection to using Levenshtein distances is that there is little apparent difference in phonological confusability between a pair like 'cats' and 'bird', which has a Levenshtein distance of 4, and a pair like 'cats' and 'pita,' which has a Levenshtein distance of only 3 but which is arguably even more different since it differs in syllable structure. Ultimately, neither pair is especially confusable: the effects of phonological confusability tail off after 1 or 2 edits.

	Dutch	English	French	German
real	4.95	4.96	5.32	5.53
μ (simulated)	4.97	4.97	5.34	5.57
σ (simulated)	0.005	0.002	0.002	0.005
z	-3.80	-6.0	-6.2	-6.9
p	**<.001**	**<.001**	**<.001**	**<.001**

Table 4.2: z- statistics comparing the average Levenshtein distance in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of average Levenshtein distance in the 30 simulated lexicon for each language.

4-5, the real lexicon has a smaller average Levenshtein distance than the simulated ones consistently across words of any length and of any language.

4.2.3 Network measures

Simply calculating phonological neighbors, however, does not tell us everything about how wordforms are distributed across a lexicon. Perhaps some words have many neighbors while others have few. Or it could be the case that neighbor pairs tend to be more uniformly distributed across the lexicon. To answer these questions, we constructed a phonological neighborhood network as in Arbesman, Strogatz, and Vitevitch (2010), whereby we built a graph in which each word is a node and any phonological neighbors are connected by an edge, as in the toy example in Figure 4-6, that shows the situation for a lexicon of 14 words.

Figure 4-7 shows examples of such networks for English 4-phone words, where each word is a node, with an edge drawn between any two words that are phonological neighbors (1 edit away). Words with no or few neighbors tend to be clustered on the outside. (The ring of points around the perimeter of the circle represent the isolates–words with no neighbors.) Words with many neighbors are, in general, plotted more centrally. We compared the shape of lexicons generated by different models to the real lexicon. As can be seen in Figure 4-7, of all the models, the 5-phone model most closely resembles the real lexicon. Substantially more clustering is observed in

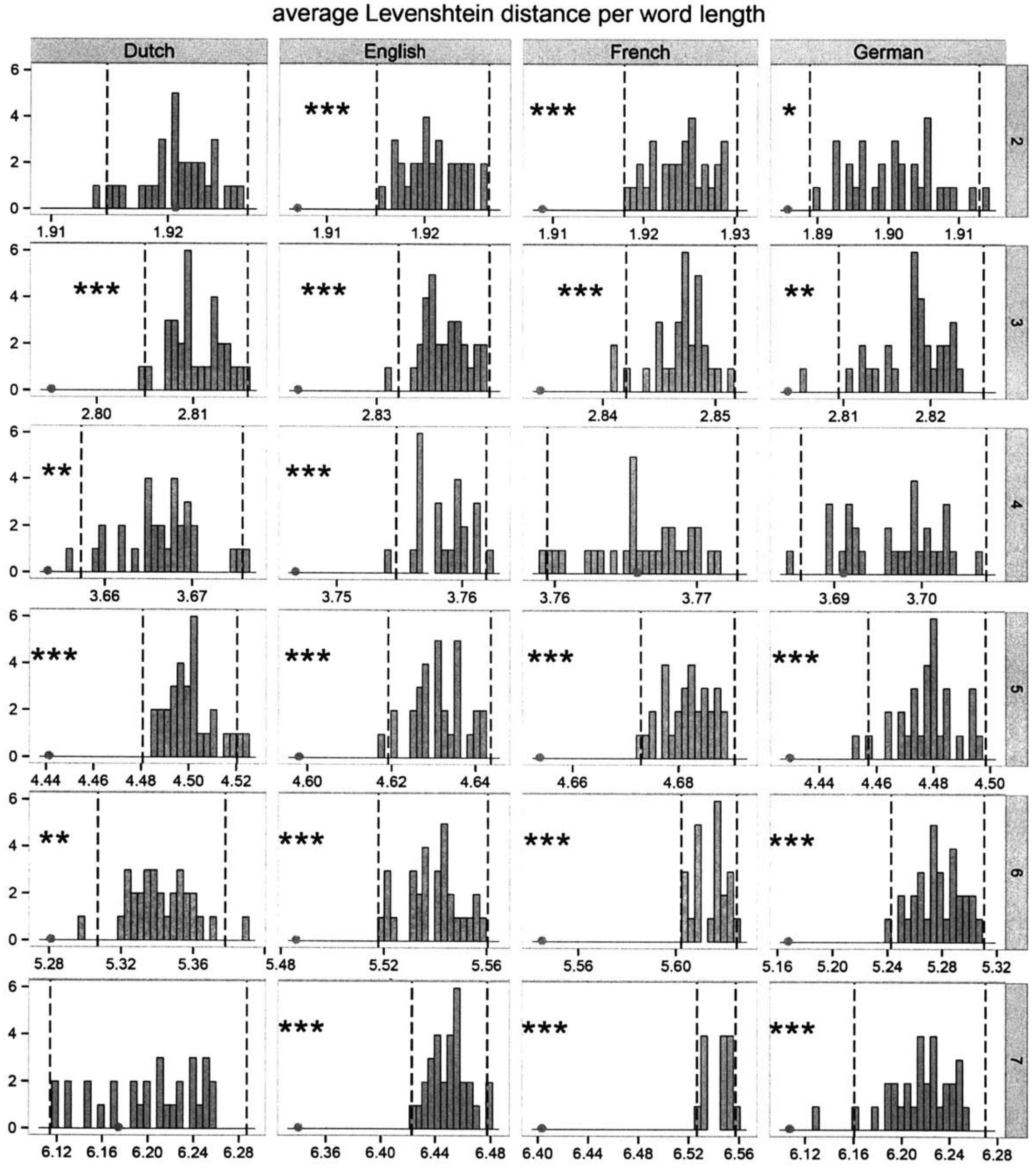


Figure 4-5: Average Levenshtein distance by word length (2-7) for each language (red dots) compared to the distribution of average Levenshtein distance obtained across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. One star represents $p < .05$, two stars $p < .01$, and three stars $p < .001$.

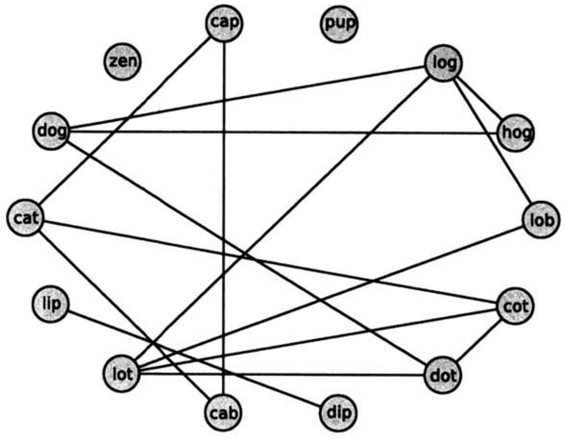


Figure 4-6: Example phonological network. Each word is a node, and any words that are 1 edit apart are connected by an edge.

the more restrictive generative models: the 5-phone, 2-syllable and PCFG models have many more connected neighbors than a 1-phone model. This corresponds to the fact that many more words are possible in the 1-phone model (e.g. 'cktw' is a possible word), than in a more constrained model that respects phonotactics. Therefore the space is largest in the 1-phone model, and the probability of generating a word that is a neighbor of a previously generated word is correspondingly lower. Crucially, however, the real lexicon seems even clumpier overall than the lexicons produced by any of the generative models.

Using techniques from network analysis that have been fruitfully applied to describe social networks and other complex systems (Wasserman & Faust, 1994; Watts & Strogatz, 1998; Barabási & Albert, 1999), we can quantitatively characterize the clustering behavior of the lexicon. We computed the *transitivity*, *average clustering coefficient*, and the percent of nodes in the *giant component*. All three of these measures can be used to evaluate how tightly clustered the words in the lexicon are. A graph's *transitivity* is the ratio of the number of triangles (a set of 3 nodes in which each node

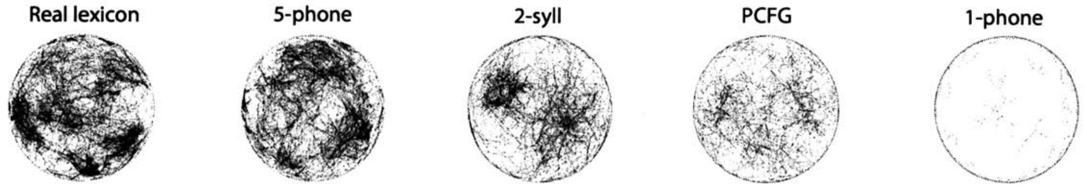


Figure 4-7: Sampling of phonological neighbor network from the different generative models applied on all 4-phone wordforms of the English lexicon. Each point is a word, and any two connected words are phonological neighbors. The simulated lexicons from less constrained generative models are less clustered and have more isolates (words with no neighbors, plotted on the outside ring).

in the set is connected to both other nodes in the set) to the number of triads (a set of 3 nodes in which at least two of the nodes are connected). Thus, transitivity in effect asks, given that A is connected to B and B is connected to C, how likely is it that A is also connected to C? The *average clustering coefficient* is a closely related measure that finds the average clustering coefficient across all nodes, where the clustering coefficient of a node is defined as the fraction of possible triangles that *could* go through that node that actually do go through that node. Both values measure the extent to which nodes cluster together. The largest cluster in a network is known as the *giant component.* A network with many isolated nodes will have a relatively small giant component, whereas one in which nodes are tightly clustered will have a large giant component. These measures give us some insight into the internal structure of the lexicon, over and above those obtained by looking at more global measures such as the number of minimal pairs and the average Levenshtein distance. If the real lexicon is clumpier than expected by chance, we predict that, relative to the simulated lexicons, the real lexicons will show higher transitivity, higher average clustering coefficients, and a larger proportion of words in the giant component.

As observed in Figure 4-8, there is no systematic difference between the real lexicon and the simulated ones regarding the average clustering coefficient measures and the percentage of nodes in the giant component. Yet there is a significant effect of transitivity (see Table 4.3). The reason that average clustering coefficient shows less of an effect than transitivity is likely that average clustering coefficient is more dependent

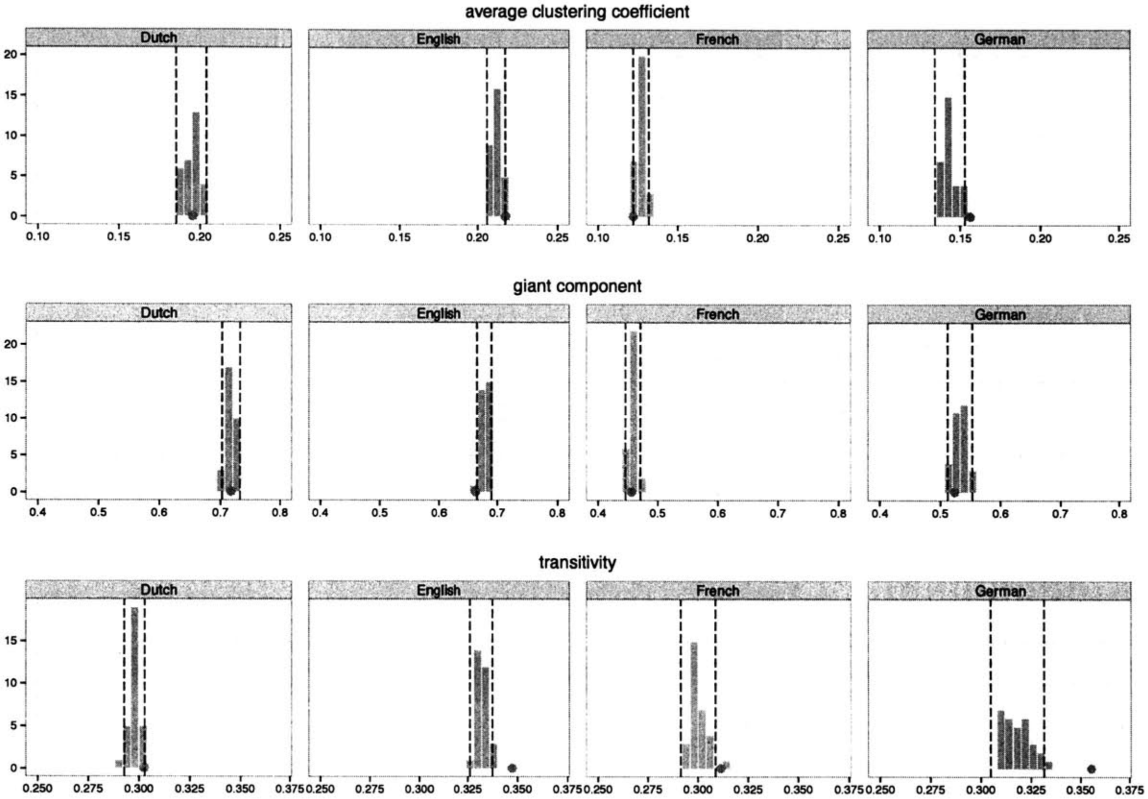


Figure 4-8: Distributions of our best generative model (the histograms) compared to the real lexicon (the red dot) in terms of network measures for lexical networks (where each node is a word and any 2 nodes that are minimal pairs are joined in the network): the percent of nodes in the average clustering coefficient, giant component, and transitivity.

		Dutch	English	French	German
Average Clustering coefficient	real	0.2	0.22	0.12	0.16
	μ (simulated)	0.19	0.21	0.13	0.14
	σ (simulated)	0.005	0.003	0.002	0.005
	z	0.1	2	-2	2.7
	p	0.9	.05	.05	**<.01**
Giant component	real	0.72	0.66	0.46	0.52
	μ (simulated)	0.72	0.68	0.46	0.53
	σ (simulated)	0.008	0.006	0.006	0.01
	z	-0.1	-2.4	-0.4	-0.9
	p	0.9	<.05	0.7	0.4
Transitivity	real	0.3	0.35	0.31	0.36
	μ (simulated)	0.3	0.33	0.3	0.32
	σ (simulated)	0.003	0.003	0.004	0.007
	z	1.8	5.4	2.6	5.5
	p	0.07	**<.001**	**<.05**	**<.001**

Table 4.3: z- statistics comparing various network measure (Average clustering coefficient, proportion of words in the giant component, transitivity) in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of these measures in the 30 simulated lexicon for each language.

on low-degree nodes, like the many isolates that exist for longer words in lexical networks (Sporns, 2011). The lack of effect for the giant component measure may simply be because the proportion of words in the giant component is not a particularly robust measure since it can be dramatically shifted by the addition or subtraction of one or two key neighbors. The higher transitivity, however, suggests that in addition to having more overall neighbors in the real lexicons, the neighborhoods themselves are more well-connected than the neighborhoods in simulated lexicons are. That is, if two words A and B are both neighbors of word C, A and B are themselves more likely to be neighbors in the real lexicon than they are in the simulated lexicons.

4.2.4 Robustness of the results

We chose as our baseline a 5-phone model because it performed best on the cross-validation test. Yet, it is important to note that any pattern of clumpiness or dispersion that we find should occur independently of this specific lexical generation model. To check whether our results were robust across the different measures of wordform similarity, we compared the same measures (minimal pairs count, average Levenshtein distance and network measures) obtained in the 3 best models according to our evaluation (see Figure 4-1): the 5-phone model, the 6-phone model and the 4-phone model.

As shown in Figure 4-9, we find qualitatively similar results with the 3 best models across all the measures of wordform similarity previously introduced.[12] In general, there were more minimal pairs and lower average Levenshtein distance in the real lexicons than across the three best models. As for the 5-phone model, no conclusive results were obtained for the average clustering coefficient and the giant component measures but the transitivity was higher in the real lexicons than in the three best models of lexicons.

This is evidence that the pattern of clumpiness we found with the 5-phone model

[12] The 3-phone model behaves somewhat differently and in fact shows more clustering than the 5-phone model. But, because its performance on the held-out data set is poor compared to the models shown here, we do not focus on this model.

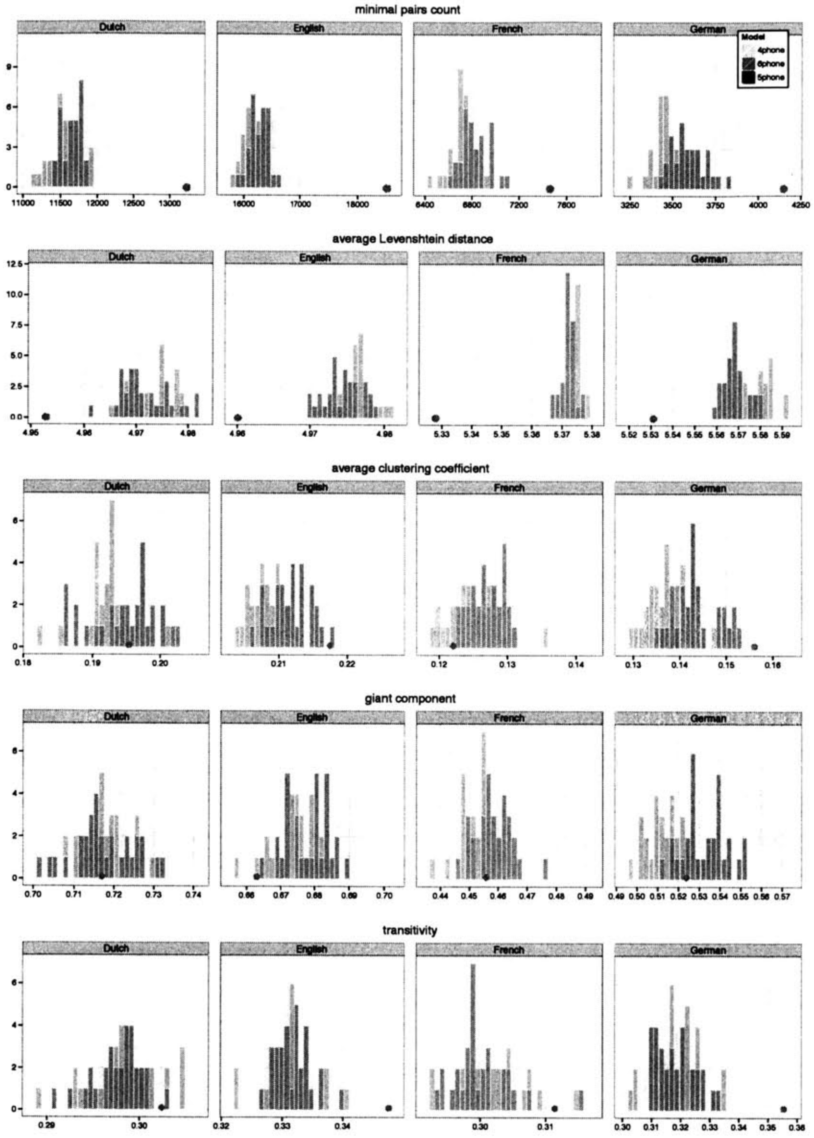


Figure 4-9: Distributions of a given measure for our best model of word generation (5-phone in dark color), our second best model (6-phone in light color) and our third best model (4-phone in translucent color) compared to the measure in the real lexicons (the red dots) for the four languages and all the measures reviewed so far.

is robust across lexical generation models. A pressure for clumpiness is thus visible beyond the particular model of phonotactic probability adopted by the best models produced here.

We also tested whether the German, Dutch, and French infinitival verb endings could be driving clumpiness effects by redoing the analyses above using just root forms (i.e., by removing the infinitival ending from the verbs). One might imagine that, because most verbs end in *-er* in French, for instance, these words have fewer degrees of freedom and thus edit distances will be smaller across the lexicon. In our analysis using just root forms, however, the results were qualitatively the same as when we used lemmas in their infinitive form, likely because the generative models already capture this regularity. That is, our baseline models too have a disproportionate number of words ending in *-er* in French and *-en* in German and Dutch. Because the presence of these infinitival stems does not substantially alter the result, we chose to keep them in the main analysis so as to be consistent with the standard databases we used (CELEX and Lexique).

4.2.5 Interim summary

In general, these measures suggest that the lexicon is clumpy: words tend to be more phonologically similar to each other than would be expected by chance. Word pairs in the real lexicon are more likely to be minimal pairs and more likely to have a small edit distance compared to words in the "chance" simulated lexicons. The chance rate here was determined through *a priori* model comparison of different plausible generating models for words. This technique has allowed us to test for clumpiness vs. dispersion while still respecting the major phonotactic tendencies in each language. It is important to emphasize that these results were computed on monomorphemes, so the results are not an artifact of morphology.

Crucially, the lexicon shows a tendency towards clumpiness above and beyond phonotactics. As we discussed earlier, phonotactic rules themselves can be thought of as a major source of clumpiness in the lexicon, insofar as phonotactics dramatically restricts the space of possible words. Yet, while our best lexical model controls for phonotactic regularity, we still observe clumpiness in the real lexicon compared to this baseline. This suggests additional pressure for clustering beyond just phonotactics.

4.3 Results: Finer-grained patterns of similarity in the lexicon

Across a variety of measures, we found that wordforms tend to be more similar than expected by chance across all languages under study. Yet, while wordform similarity might be explained by a variety of cognitive advantages (see Introduction), it does not necessarily follow that the lexicon is not subject to communicative pressure favoring wordform distinctiveness. A possibility is that the similarity between wordforms may not be uniformly distributed across the real lexicon but may be constrained by other dimensions that maximize their distinctiveness in the course of lexical processing, such as:

1. **phonological distinctiveness:** Not every pair of phonemes is equally confusable. For instance, a minimal pair like 'cap' and 'map' are unlikely to be confused since /k/ and /m/ are quite distinct. But 'cap' and 'gap' differ by only the voicing of the first consonant and are thus much more confusable (Miller & Nicely, 1955). From a communicative perspective, this more subtle contrast is potentially much more troublesome for communication and is therefore more likely to be avoided. So even though the number of minimal pairs is higher than expected by chance in natural lexicons, this might not be problematic for communication as long as they are not based on confusable contrasts.

2. **grammatical categories:** Not every pair of words is equally confusable. For instance, nouns (e.g. 'berry') are more likely to be confused with other nouns (e.g., 'cherry') than words from another grammatical category (e.g., the adverb 'very') because they appear in a noun syntactic context which constrains listeners to expect a noun in this position. Therefore, from a communicative point of view,

there should be more minimal pairs distributed across syntactic categories than within the same syntactic category to minimize the risk of miscommunication.

In the following we test how the simulated lexicons compare to the real lexicons along these two dimensions.

4.3.1 Wordform distinctiveness in minimal pairs

The accurate recognition of a word depends on the distinctiveness of the phonological contrasts distinguishing words. If lexicons aim to minimize confusability, they should prefer distinctive contrast minimal pairs as opposed to confusable ones. In the case of 'cap' and 'map,' for instance, one word is unlikely to be confused for the other since the contrast is quite distinctive. But one retains the benefits of being able to re-use the word coda *-ap* in both cases. Thus, it is possible that lexicons can have the learning benefit of having frequent minimal pairs, as long as they are not based on confusable contrasts.

To evaluate this hypothesis, we looked at the 5% most frequent minimal pair contrasts and derived a measure of confusability for these contrasts. Phonemes can be characterized by their phonological features: place of articulation (e.g., labial, dental, palatal), manner of articulation (e.g., stop, fricative, glides) and voice for consonants (voiced, unvoiced); height (close to open), backness (front to back) and roundness for vowels. For each of the 5% most frequent pairs of contrasts, we calculated the difference in phonological features between each member of the pair. For example the pair /k/ and /m/ has 3 features that differ: place, manner and voicing. The test statistic that we use here is the average number of features that differ in a minimal pair. This measure ranges from 1 (highly confusable) to 3 (highly distinguishable).[13]

Figure 4-10 shows the average number of features that differ in the 5% most frequent minimal pair contrasts in the real lexicon and across all simulated lexicons for each language. The minimal pairs contrasts in the real lexicon are no more distinguishable in phonetic space than are the minimal pairs in the chance lexicon. This indicates

[13] For French we added nasalization as a vowel feature. The measure for French vowel contrasts therefore ranged from 1 to 4.

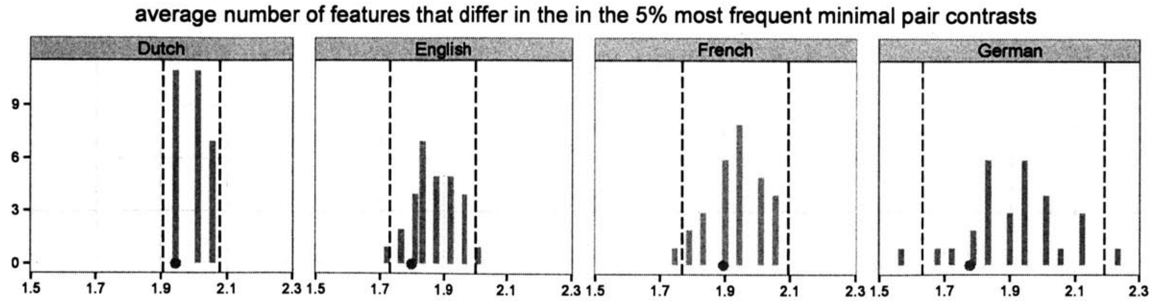


Figure 4-10: Distributions of the average number of feature difference for the 5% most frequent minimal pair contrasts in the simulated lexicon compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. There is no evidence that these frequent contrasts are more perceptible than expected by chance (all $ps > .30$).

that minimal pairs do not rely on more perceptible contrasts for distinctiveness than what is expected by phonotactics alone.

The previous measure showed that frequent minimal pair contrasts are not more perceptible than expected by chance alone. But because we used a coarse measure of confusability (the average number of different phonological features) looking only at the most used contrasts, it could still be the case that a more perceptual and language-specific measure of phoneme confusability–looking at a broader range of possible contrasts–would be a better predictor of clumpiness. If the lexicons prefer minimal pairs to be distinctive then we should observe more minimal pairs having easily perceptible contrasts than those having confusable contrasts. In order to investigate this possibility, we looked at minimal pairs in English for which confusability data between phonemes are readily available (Miller & Nicely, 1955). We computed the distance between the mean number of minimal pairs in our simulated lexicons and the number of minimal pairs in the real lexicon for each of the 120 contrasts present in the Miller and Nicely dataset. The distance is simply the difference between a) the mean number of minimal pairs in the simulated lexicons and b) the number of minimal pairs in the real lexicon, divided by the standard deviation of the value across the 30 simulated lexicons. In effect, this acts as a z-score that tells us how far the real lexicon value falls from what we expect under a null model.

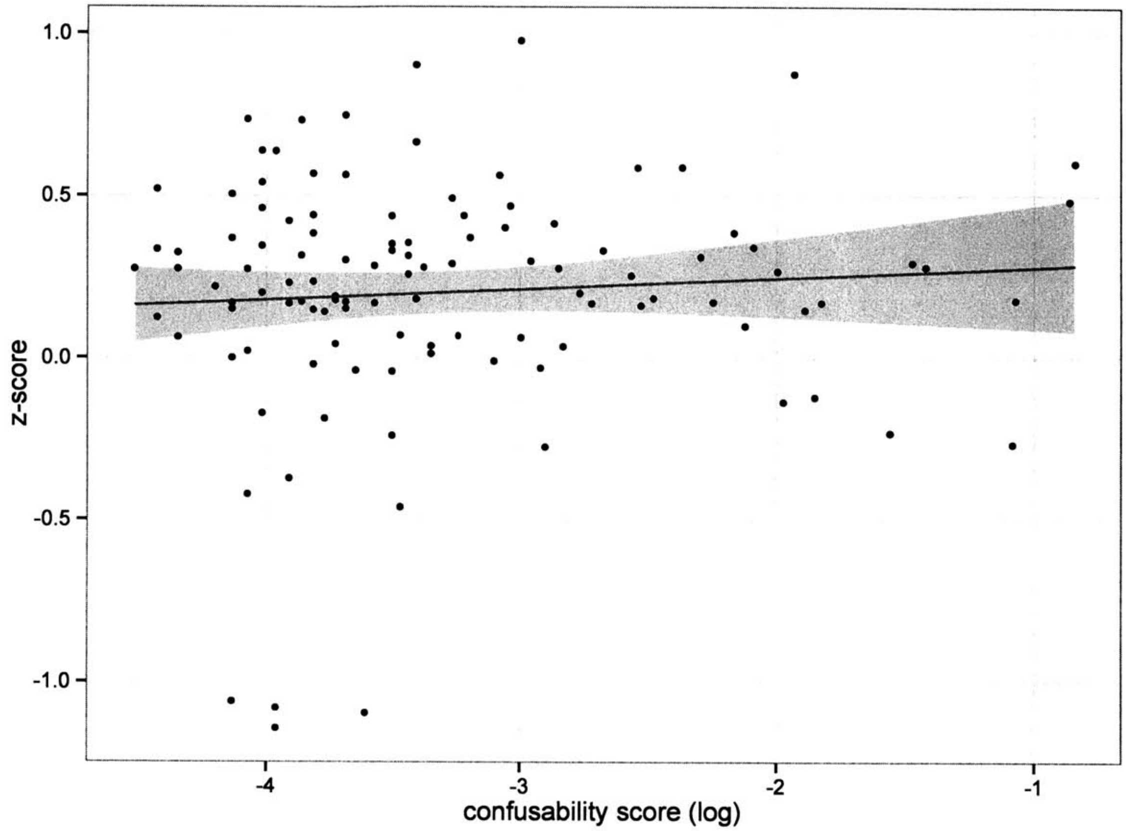


Figure 4-11: z-score obtained between the mean number of minimal pairs in the real lexicon and in the simulated lexicons for each of the minimal pair contrasts present in the Miller and Nicely's dataset as a function of their (log) confusability.

Figure 4-11 shows the z-score obtained for each phonemic contrast as a function of its confusability (the higher the more confusable). As can be observed, there is no effect of confusability on the z-score ($p > 0.5$). That is, there is no evidence that the English lexicon is more clumpy around highly distinctive contrasts than around highly confusable contrasts.

Thus, it appears that the clumpiness effect is driven not just by highly distinct sound sequences but is present even when considering highly confusable sounds. This points to a pressure for lexical clumpiness which may work against robust communication.

4.3.2 Wordform similarities within and across grammatical categories

Words do not usually appear in isolation but are embedded in richer linguistic context. A wealth of studies show that adults and children use the context of a sentence to constrain lexical access (Borovsky, Elman, & Fernald, 2012; Altmann & Kamide, 1999). Hence even if the lexicon is clumpy as a whole, the context might be sufficient to disambiguate between two similar wordforms. One obvious contextual disambiguation is the syntactic category of words. For example, consider the sentence "did you see my sock?" The chance that a native English speaker might confuse the word 'sock' with 'lock' in the context of following 'my' might be greater than confusing 'sock' with 'mock', because 'lock' is a noun–which is consistent with the syntactic context–whereas 'mock' is a verb, which is inconsistent with the syntactic context. Moreover, because children as young as 18-months have been shown to use function words to recognize and learn the difference between verbs and nouns on-line, these sorts of categorizing effects may be crucial to language acquisition (Cauvet et al., 2014).

As with the lexicon more broadly, there are two possible outcomes that could arise from comparing word forms within as opposed to across syntactic categories. On the one hand, because context is usually enough to distinguish among different parts of speech, confusability of words should be less of a problem across syntactic categories. That is, even though 'bee' and 'see' are minimal pairs, one is unlikely to misperceive "I was just stung by a *bee*" as "I was just stung by a *see*." This account predicts more similarity across syntactic categories than within syntactic categories. On the other hand, the effects of learnability and ease of processing may be enhanced by having increased similarity between words of the same part of speech. That is, having nouns that sound like other nouns and verbs that sound like other verbs could convey a processing advantage. Under this account, we would expect more similarity within as opposed to between syntactic category.

For this evaluation, we used the Part Of Speech (POS) tags in CELEX for Dutch, English and German and in Lexique for French to count the number of minimal pairs

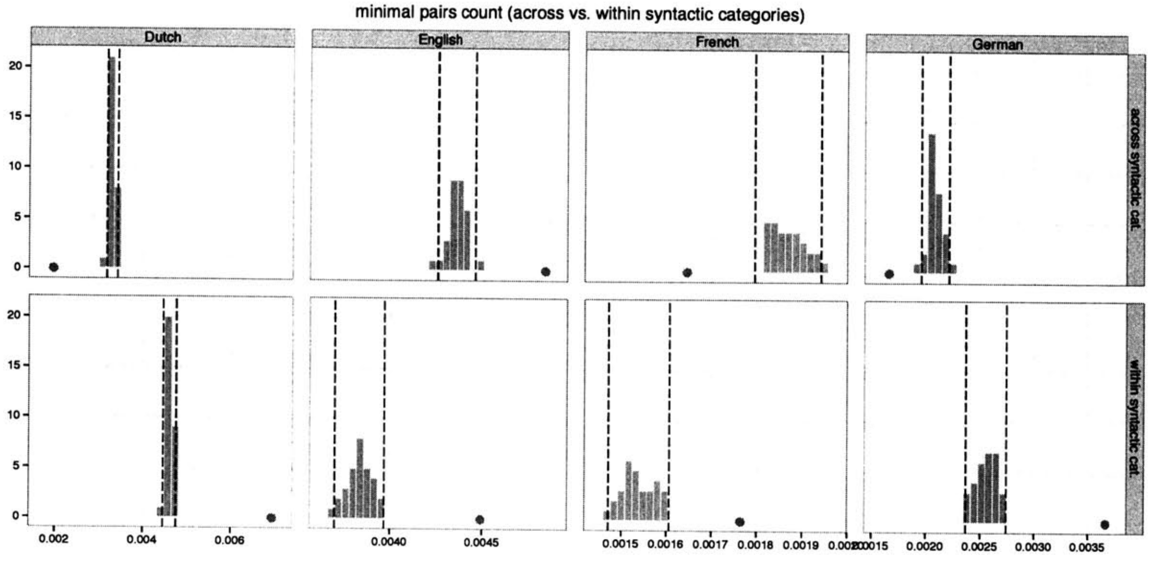


Figure 4-12: Distributions of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. All 4 languages are significantly more likely to have minimal pairs within categories than would be expected by chance.

within the same syntactic categories (e.g., 'lock' / 'sock') and across different syntactic categories (e.g., 'mock' / 'sock'). For each simulated lexicon, we randomly assigned the syntactic categories of real words of length n to generated words of length n and similarly counted the number of minimal pairs appearing within and across the same syntactic categories.[14] Note that for wordforms having several syntactic categories in the real lexicon (homophones, e.g.,'seam'/'seem' which are counted as a single wordform in our lexicons, /sim/), we chose the syntactic category of the most frequent items (e.g., because the most frequent meaning of /sim/ is 'seem' it will be categorized as a verb). Because there are more across-category minimal pairs than within-category minimal pairs across languages, we divided the number of minimal pairs appearing across and within categories by the number of across- and within-category word pairs respectively. The final measure is thus the probability of getting a minimal pair, across categories or within categories.

As before, we compare the real lexicon to the simulated lexicons but break the

[14] This was to ensure that certain categories, such as pronouns, which are reserved for smaller words will not be assigned to longer words.

		Dutch	English	French	German
across syntactic categories	real	0.002	0.0048	0.0016	0.0017
	μ (simulated)	0.0033	0.0044	0.0019	0.0021
	σ (simulated)	1e-04	1e-04	1e-05	1e-04
	z	-21.5	9	-6	-6.6
	p	**<.001**	<.001	**<.001**	**<.001**
within syntactic categories	real	0.0069	0.0045	0.0018	0.0037
	μ (simulated)	0.0046	0.0038	0.0015	0.0026
	σ (simulated)	1e-04	1e-04	1e-05	1e-04
	z	31.2	9.6	6.5	11.7
	p	<.001	<.001	<.001	<.001

Table 4.4: z- statistics comparing the probability of getting a minimal pair within and across syntactic categories in the real lexicon with the chance distribution of mean μ and standard deviation σ corresponding to the distribution of the probability of having a minimal pair in the 30 simulated lexicon for each language. The red p-values shows a significant effect of clumpiness and the blue ones a significant effect in the opposite direction.

measures down by similarity within syntactic category (only looking at the similarity of nouns to other nouns, verbs to other verbs, and so on) and between syntactic category (only looking at the similarity of nouns to non-nouns, verbs to non-verbs, etc.). As shown in Figure 4-12, we found that there are *more* minimal pairs within the same syntactic category in the real lexicons than would be expected by chance for all 4 languages. That is, for within syntactic category analyses, all four languages are clumpier than expected under the null models. For the across-category analysis, the result is less clear. For French, German, Dutch, there are *fewer* minimal pairs across different syntactic categories than would be expected by chance. For English, there are more across-category minimal pairs than expected by chance.

A subsequent post-hoc analysis found that the unclear results for the across-category analysis can in part be explained by the infinitival affixes that appear on French, Dutch, and German verbs. When we remove these verb endings, the across-category differences look roughly like what one expects by chance (see Figure 4-13). This result is unsurprising since the presence of verb stems like -*er* means that any given verb is less likely to be a neighbor of a noun since most nouns do *not* end in -*er*.

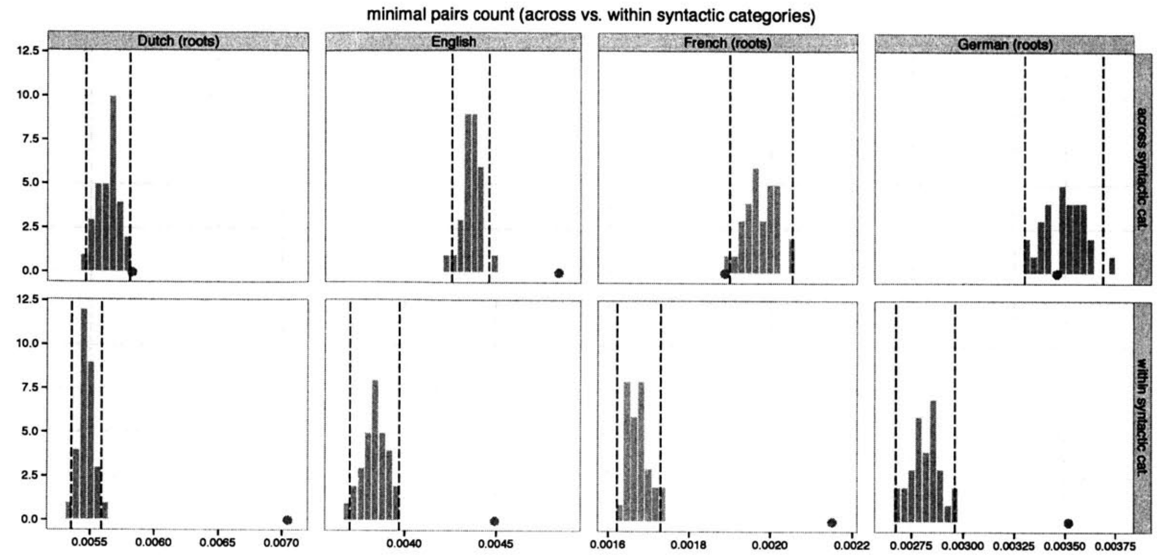


Figure 4-13: As in Figure 4-12, these histograms show the distribution of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon, but without infinitive endings on verbs in Dutch, French and German.

The within-category analysis is qualitatively unchanged by focusing on roots (in all cases the real lexicon is clumpier than expected by chance).

Note that the probability of getting a minimal pair within the same syntactic category is greater than the probability of getting a minimal pair across different syntactic categories for Dutch, French and German but not for English. A possible explanation for this difference is that there is still some verbal morphology present in the lemmas for Dutch, French and German that we could not capture, and this morphology artificially inflates the number of within-category minimal pairs compared to the number of across-category minimal pairs. For instance, in Dutch, verbs of motion systematically display phonaesthemes (typically a schwa followed by a sonorant) that are not analyzed as suffixes. Another possibility for this difference is that the probability of getting a minimal pair across and within syntactic categories may not be directly comparable because the length distributions for within category words and across categories words are different and may thus drive part of the difference found here. As a result we prefer to concentrate on the comparison of the real lexicon with the simulated lexicons.

4.3.3 Interim summary

To sum up, we did not find evidence that clumpiness is more likely among perceptible than confusable phonological contrasts. That is, it seems that confusable phoneme pairs like 'm' and 'p' are just as likely to be the basis of minimal pairs as less confusable pairs. One possible explanation for this null result is that even highly confusable phoneme pairs like 'b' and 'p' are only confusable in certain specific contexts, such as after vowels at the end of words as in 'cab' and 'cap' (Steriade, 1997). Even then, though, context might be enough to disambiguate the words such that the confusability is not an issue.

We found evidence for more clumpiness within syntactic category than across syntactic categories. This may potentially be the consequence of a more general pattern: words of the same syntactic category may share more phonological properties than with words of different classes (Kelly, 1992). For English words, it is also the case that we see more clustering across categories than expected by chance. But that is not the case for French, German, or Dutch when we control for the presence of infinitival markers. Therefore, at least for these languages, it may even be the case that this syntactic category effect drives the larger clumpiness effect observed across the lexicon. This would be consistent with the findings of Monaghan et al. (2014) and Dautriche et al. (submitted), who show a relationship between semantic and phonological similarity.

4.4 Discussion

We have shown that lexicons use their degrees of freedom in a systematic and interesting way. While we can still characterize the relationship between wordforms and meanings as arbitrary, structure emerges when one considers the relationships within the space of possible wordforms. Across a wide variety of measures of phonological similarity, the real lexicons of natural languages show significantly more clustering than lexicons produced by the "best" generative model selected by our held-out model comparison procedure.

Because we focused on monomorphemic words, this effect cannot be a result of words sharing prefixes and suffixes. It is also not a product of any structure captured by sound-to-sound transition probabilities such as phonotactic regularities, since our models capture these patterns. This last point is crucial: even though our model took away some clustering effect by capturing sound-to-sound transition probabilities (compare the density of neighborhood between the network of the 1-phone model to the 5-phone model in Figure 4-7), there is still some clustering effect that is not explained by frequency distribution of groups of phonemes.

Certainly, one explanation for the clumpiness in the lexicon is shared phonetic properties of semantically related words. Like 'skirt' and 'shirt', many words in the language share deep etymological roots. Moreover, the presence of sound symbolism in the lexicon is another source of structure in the lexicon not captured by our models. For instance, there is a tendency in English for *gl-* words to be associated with light reflectance as in 'glimmer', 'gleam' and 'glisten' (Bloomfield, 1933; Bergen, 2004). There are additionally cross-linguistic correspondences between form and meaning, such as a tendency for words referring to smallness to contain high vowels (Hinton, Nichols, & Ohala, 2006; Sapir, 1929). Interestingly, recent studies show that phonologically similar words tend to be more semantically similar across measures of wordform similarity over the whole English lexicon (Monaghan et al., 2014) but also in Dutch, French and German. This suggests that clumpiness in the lexicon cannot be attributed to small islands of sound symbolism. Rather, it reveals a fundamental drive for regularity in the lexicon, a drive that conflicts with the pressure for words to be as phonologically distinct as possible. Importantly,

One possible source of the lexicon's clumpiness is that speakers preferentially re-use common articulatory sequences. That is, beyond just phonotactics and physical constraints, speakers find it easier to articulate sounds that they already know. Recall our example of the language in which there is only one word for a speaker to learn. She would quickly become an expert. Along those lines, the presence of any given sound sequence in the language makes it more likely that the sequence will be re-used in a new word or a new pronunciation of an existing word. In that sense, the lexicon 'overfits': any new word is deeply dependent on the existing words in the lexicon. Note that because our baseline used a lexical generation model, any pressure for re-use must occur over and above the observed statistical trends (e.g., 5-phone sequences) in the language.

Relatedly, lexical clumpiness may be advantageous for some aspects of word production. While words having many neighbors are challenging for word recognition (Luce, 1986; Luce & Pisoni, 1998), they may be easy words to produce (Gahl et al., 2012; Vitevitch, 2002; Vitevitch & Sommers, 2003). Previous studies suggest that listener-oriented model of speech production– where speakers adjust their speech to ensure intelligibility of words that might otherwise be difficult to understand (as could be words with many neighbors)– are limited by attentional demands and working memory in conversational speech (Arnold, 2008; Lane, Groisman, & Ferreira, 2006). However, speakers may produce words with many neighbors faster, because they are easier to access and retrieve (Gahl et al., 2012; Dell & Gordon, 2003). Hence a clumpy lexicon would be beneficial for a speaker-oriented model of speech production associated with rapid lexical access and retrieval.

A clumpy lexicon also may allow for easier compression of lexical knowledge. By having words that share many parts, it may be possible to store words more easily. Though we concentrate here on monomorphemic lemmas, these account only for one third of all the lemmas in the lexicon. The fact that languages reuse words or parts of words in the remaining two thirds of the lemmas shows that re-use of existing phonological material must be important. It may even be the case that, much as morphology allows the productive combination of word parts into novel words, there exist sound sequences below the level of the morpheme that *also* act as productive units of sound.

Phonological proximity may display also some functional advantages in the context of word learning. To form a novel lexical entry in their lexicon, children must be able to extract a word form and associate it to a meaning. In theory, a clumpy lexicon may be advantageous for learning as it reduces the amount of new information that must be represented in the lexicon. For instance, to learn a novel word such as 'blick', children need to create a novel phonological representation that needs to be associated to a novel semantic representation. Re-using parts of existing phonological forms may be more efficient because it allows children to minimize the amount of phonological information that must be learnt and remembered (Storkel, Maekawa, & Aschenbrenner, 2012; Storkel & Maekawa, 2005).

Despite the fact that one might expect the lexicon to be maximally dispersed for communicative efficiency, these results strongly suggest that the lexicon is not nearly as sparse as it could be–even given various phonetic constraints. Thus, why does communicative efficiency not conflict with clumpiness in the lexicon?

One possibility is that clumpiness does not appear randomly in the lexicon but is organized along dimensions that maximize wordform recoverability. We hypothesized that recoverability could be enhanced if similar wordforms such as minimal pairs were disambiguated by minimally confusable sounds. Our results provide no evidence that the lexicon is less clumpy for confusable sounds than for non-confusable sounds. Relatedly, lexical access might be faster in a lexicon where confusable wordforms span different syntactic categories. Yet we find that, if anything, wordforms are more similar *within* the same syntactic category than what would be expected by chance for all four languages despite the absence of morphology.

Another possibility that would explain why communicative efficiency does not conflict with clumpiness in the lexicon is that contextual information outside the word pronunciation is usually enough to disambiguate words. Therefore it simply does not matter whether certain words are closer together in phonetic space than they might otherwise be. Piantadosi et al. (2012) showed that lexical ambiguity, such as dozens of meanings for short words like *run*, does not impede communication and in fact promotes it by allowing the re-use of short words. In a similar way, there may be a communicative advantage from having not just identical words re-used but from re-using words that are merely similar. In all cases, context may be enough to disambiguate the intended meaning and avoid confusion–whether it be confusion between two competing meanings for the same word or confusion between two similar-sounding words.

Likewise, our analysis here concentrated on the phoneme representation of words ignoring the fact that speech contains a lot of fine phonetic details that listeners could use to disambiguate between words. For instance, pairs of homophones such as 'thyme'/'time' in English can be differentiated based on their duration (Gahl, 2008). Kemps, Wurm, Ernestus, Schreuder, and Baayen (2005) show that English and Dutch listeners are sensitive to fine-grained durational differences between a base word ('run') and the base word as it occurs in an inflected or derived word ('runner'). Being sensitive to these cues may also be useful to disambiguate between words that sound similar such as minimal pairs.

Clumpy lexicons seem to be advantageous for word production, word learning and memory but detrimental for word perception. Yet the interaction of these cognitive and articulatory constraints with a pressure for clumpiness or a pressure for dispersion is complex. Clearly there are many functional pressures that are at play for the listener, the speaker and the learner, and they do not individually point towards either clumpiness or distinctiveness of wordforms. In the context of word learning, wordform similarity may be both advantageous and disadvantageous: Similar-sounding words minimize the amount of information that needs to be stored (e.g., Storkel & Maekawa, 2005), help in word segmentation (Altvater-Mackensen & Mani, 2013), are easier to recognize because composed of highly-probable sequences of sounds (e.g., Jusczyk & Luce, 1994) and help children group words into categories (i.e., nouns, verbs) when phonological proximity is aligned with semantic or syntactic classes (Monaghan et al., 2011). Yet when it comes to individual word learning, learners have a hard time to associate a novel meaning to word that sound close to one they know (e.g., 'tog' a phonological neighbor of the familiar word 'dog'; e.g., (Swingley & Aslin, 2007)) and this disadvantage is even greater when phonological similarity is aligned with syntactic or semantic similarity (Dautriche et al., 2015). Similarly in the context of word production: while clumpy lexicons may be easier to produce, they may also give rise to a greater number of speech errors when the relationship between phonological proximity and semantic proximity is high (Dell & Reich, 1981). Importantly, our results suggest that the functional challenges associated with wordform similarity are weighted less than its functional advantages. In other words, the sum of all these functional pressures (for the listener, the speaker, the learner) pushes towards a clumpy lexicon.

Certainly, while we can probably get an idea of the weight of different functional pressures from observing the structure of the lexicon (and of languages more generally), we cannot tell whether they actually explain why lexicons look the way they do. Work looking at language evolution offers a promising venue to understand how functional pressures from both language usage and language learning combine to produce the particular pattern of clumpiness observed in human languages. By observing how language is transmitted culturally from one generation to the next, either using computational models or experiments with human participants in the lab, it is possible to isolate how languages are shaped by the processes of both cross-generation transmission (language learning) and within-generation communication (language use) (Kirby, Tamariz, Cornish, & Smith, 2015; Kirby & Hurford, 2002; K. Smith et al., 2003; Kirby, Cornish, & Smith, 2008). The methodology used here, whereby the real lexicon is compared to a distribution of statistically plausible 'null' lexicons, could be used to generate hypotheses about the lexicon and human language more generally that could be tested experimentally using such language evolution techniques. While much previous work has focused on simply measuring statistical properties of natural language, modern computing power makes it possible to simulate thousands of different languages with different constraints, structures, and biases. By comparing real natural language to a range of simulated possibilities, it is possible to assess which aspects of natural language occur by chance and which exist for a reason.

Of course, we must keep in mind that the present work examines only a small number of European languages. To know whether the effect generalizes would require a larger number of languages, and we undertake exactly such a project in other works (Dautriche et al., submitted; Mahowald, Dautriche, Piantadosi, & Gibson, Under revision). Specifically, we use a corpus of 100+ languages from Wikipedia to show large-scale evidence for a) more frequent words to be more orthographically probable and have more minimal pairs than less frequent words and b) for semantically related words to be more phonetically similar than less related words. While the Wikipedia corpus does not focus on monomorphemes and is therefore less controlled than the results presented here, it suggests that the clumpiness we observe in the lexicons of Dutch, English, German, and French likely generalizes to other languages as well.

In future work, it may be possible to test increasingly sophisticated models of phonotactics using this methodology. One possibility is that our models of phonotactics are simply not good enough yet to capture the rich structure of natural language. But the results here suggest that any "null" model that can approximate natural languages will need to account for not just the preferred sounds of a language but for the entire space of existing words. That is, the goodness of 'dax' as an English word depends not just on an underlying model of English sound structure but on the fact that 'lax' and 'wax' are words, that 'bax' is not, and on countless other properties of the existing lexicon. Another possibility is that our models of phonotactics capture more than the phonotactic constraints of languages (see above for others possible sources of clumpiness). It would be thus informative for future work to separate clearly phonotactic constraints from other sources of regularity to have a more thorough picture on how clumpiness patterns can be interpreted.

Overall, we have shown that lexicons are more richly structured than previously thought. The space of wordforms for Dutch, English, German and French is clumpier than what would be expected by the best chance model by a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. The strongest evidence comes from minimal pairs, for which the effect size was quite large. From this, we conclude that the clustered nature of the lexicon holds over and above the patterns that are captured by a phonotactic model. Underlying the pressure for dispersion in the lexical system is a deep drive for regularity and re-use beyond standard levels of lexical and morphological analysis.

Part II

Evaluating methods in psychology and linguistics

Chapter 5

Methods in linguistics: SNAP (Small N Acceptability Paradigm) for Linguistic Judgments

5.1 Introduction

We now turn to the meta-scientific part of the thesis, where I will present two studies: the first a meta-analysis of acceptability judgments in linguistics, the second a meta the extensive published literature on syntactic priming. We begin with the acceptability study.

Historically, the method in syntax and semantics research was for the researcher to use his or her own intuitions about the acceptability of phrases and sentences. This informal method worked when the field was developing, and the contrasts were large as in (1), but as the field progressed, the contrasts needed for deciding among competing theories became more complex, and the judgments consequently became more subtle, as in (2):

(1) Chomsky (1957)

a. Colorless green ideas sleep furiously.

b. Furiously sleep ideas green colorless.

(2) Chomsky (1965)

a. What do you wonder who saw?

b. I wonder what who saw.

Since the early days of generative grammar, researchers have been asking questions about methods in linguistics research, specifically the relationship between grammaticality and acceptability (Chomsky, 1986; Labov, 1978; Levelt, Van Gent, Haans, & Meijers, 1977; McCawley, 1982), the reliability of intuition as a method, and the practice of linguists using their *own* intuitions as opposed to consulting naïve speakers (Birdsong, 1989; Householder, 1965; Spencer, 1973). For a more detailed history of these issues, see Schutze (2006). In recent years, increases in the availability of data have led to further discussions of the weaknesses of the informal method (Arppe & Järvikivi, 2007; Cowart, 1997; Gibson & Fedorenko, 2013, 2010a; Gibson, Piantadosi, & Fedorenko, 2013; Gross & Culbertson, 2011; Schutze, 2006; Sorace & Keller, 2005; Wasow & Arnold, 2005; Linzen & Oseki, 2015). Such weaknesses include potential cognitive biases on the part of the researcher and participants, difficulty in controlling for discourse context, the inability to find interactions among factors, and the inability to find probabilistic effects or relative effect sizes. Furthermore, for readers who do not natively speak the language in question, it is difficult to evaluate the size of informally reported contrasts. For this reason, an advantage of formal methods is that they provide fellow researchers with quantitative information about the quality of data that is gathered: quantitative details enable an understanding of which comparisons support a theory, and which do not.

Below, we address several remaining arguments against the widespread adoption of quantitative methods in syntax and semantics research. We then present a formal replication of a large-scale experiment by Sprouse, Schütze, and Almeida (2013) on a set of sentences sampled from *Linguistic Inquiry* 2001-2010. Using these data from 100 pairwise comparisons randomly sampled from the same set of articles investigated by SSA, we present a novel proposal and provide empirical support for a Small N Acceptability Paradigm for linguistic judgments (SNAP judgments) which is robust to noise and which could dramatically decrease the burden on language researchers.

Arguments in favor of quantitative methods in syntax and semantics research

A. The current error rate in informal linguistic judgments is not as low as it could be.

SSA accept that there may be published judgments that would not be found in large-scale experiments, but they note that it is important to know the rate at which such examples occur. Consequently, SSA analyzed the judgments from 148 randomly sampled English acceptability judgments from Linguistic Inquiry (LI) 2001-2010. 127 out of these 148 experiments (86%) resulted in significant effects in the predicted direction using magnitude estimation; 130 (88%) resulted in significant predicted effects using Likert ratings; and 140 (95%) resulted in significant predicted effects in a forced choice experiment (where all values here are obtained using mixed models, which are most appropriate for this type of data (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). 7 of the 148 experiments (5%) do not show predicted effects in any of the 3 experiments. SSA generalize from the 95% rate that informal acceptability intuitions reported in research articles on generative syntax have similar statistical properties as quantitative experiments comparing acceptability ratings for various experimental items by naive participants, because each allows an error rate of approximately 5%. That is, because a 5% error rate is the acceptable standard in cognitive psychology experiments, this error rate should also be acceptable in linguistic judgments. SSA write: "The field of experimental psychology has, by consensus, signaled a willingness to tolerate a divergence of 5% over the long run between the decision to classify differences as statistically significant and whether there is a real difference between the conditions" and that, while they do not unqualifiedly endorse 5% as an acceptable error rate, they find it to be a "reasonable starting point for the current discussion."

Following Gibson, Piantadosi & Fedorenko (2013), we believe that the current error rate in published linguistic judgments could be lower without "crippl[ing] linguistic investigation" (Culicover & Jackendoff, 2010). Much of the recent debate on quantitative methods in syntax and semantics has focused on whether or not a significant p-value (p < .05) is obtained through a quantitative experiment (see Sprouse and Almeida 2012 for more discussion of effect size and statistical power in linguistic acceptability judgments). If the null hypothesis can be rejected, the syntactic judgment is said to "replicate." While SSA do not make a strong claim as to the appropriate role of Null Hypothesis Significance Testing (NHST) in syntax and semantics judgments, we believe that there *is* a place for understanding the significance of judgments in formal experiments. But we do not believe that the standards developed in the NHST paradigm are unproblematically applicable to linguistic judgments.

For one, a $p < .05$ false-positive *threshold* for NHST in behavioral experiments is not comparable to a 5% false-positive *rate* in published acceptability judgments. The NHST paradigm assumes that one has performed statistical significance testing for each particular effect under consideration; the $p < .05$ threshold is an easy way to classify the results of these tests, but it does not substitute for the important quantitative information gathered about each individual effect and the assumption that the particular sample being studied is drawn from a larger pool. On the other hand, a 5% error rate in linguistic acceptability judgments suggests that 5% of all judgments would diverge from the results of a formal experiment. But there is no sampling being done, the method provides no quantitative information about any *individual* effect. Therefore the notion of a statistical threshold across a body of judgments is problematic. If the average linguistics paper has 33 examples (the average number of US-English examples found in the papers examined by SSA), divergences are uniformly distributed, and the divergence rate is 5%, then *every* paper is likely to contain a questionable judgment: 1.64 on average.

B. Only formal experiments can give detailed information on the size of effects.

In the case of acceptability judgments, it is rarely the case that researchers actually care whether a sentence is some infinitesimally tiny amount better than another one. In fact, given a large enough sample size, one is likely to be able to find a statistically significant difference between *any* two sentence types that minimally differ. It is far more informative to investigate the size of the effect. In a rating study, the effect size can be measured as the difference in mean rating between Sentence A and Sentence B in terms of standard deviations. In a forced choice study, effect size can be estimated by the proportion of participants who choose Sentence A over Sentence B. Having a standardized system for obtaining and reporting native speaker judgments would allow readers to know just how strong the generalization in question is. That is, when we see two sentences being compared, one with a * and one without, does that mean that 51/100 people would prefer the unstarred sentence? 90/100? 100/100? To be sure, we do not think it is necessary, or even useful, to apply a uniform quantitative threshold for acceptability contrasts. Our point is simply that *some* quantitative information about the size of the effect is useful for meaningfully interpreting individual acceptability contrasts.

C. Informal linguistic experiments make it difficult for researchers who either (a) are from other fields or (b) don't speak the target language in the materials.

Even if the informal linguistic judgments in journals agreed with results from formal experiments 100% of the time, there are still important reasons for performing formal experiments. Reporting statistically valid inferences about sentence judgments would make linguistics more accessible to researchers in other fields and to researchers who are unfamiliar with the language in question. This concern may be especially relevant in light of recent evidence that acceptability judgments in non-English languages may be even lower than those in English (Linzen and Oseki 2015). Formal experiments using a consistent methodology make it easy to compare effects across languages.

D. Formal experiments need not be costly or time-consuming since even very few participants can sometimes give meaningful results.

Another common concern is that formal experiments in syntax and semantic are too costly in time and money to justify the benefits (Culicover and Jackendoff 2010). Although there is some cost to doing an experiment, the cost is now relatively low because of the existence of crowd-sourcing platforms like Amazon.com's Mechanical Turk that can provide robust results for cognitive behavioral experiments (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012; Sprouse, 2011). Such platforms provide cheap, reliable, fast labor, and there is free software available to perform such experiments (e.g., Turkolizer, as described in Gibson, Piantadosi, and Fedorenko (2011). In a syntax judgment experiment on Mechanical Turk, the researcher typically posts a survey consisting of a set of sentences (either visually or through audio) and asks for a forced choice judgment, a rating, or some other measure of acceptability. Participants fill out these surveys, and the researcher receives the data—usually within a few hours.

Still, it may seem like overkill to run a large-scale experiment to find out that "Rat cat ate the" is a less good English sentence than "The cat ate the rat." Moreover, not all researchers have easy access to Mechanical Turk or sufficient funding to run large-scale experiments. This lack of access to crowdsourcing platforms may especially affect those outside the United States (where Mechanical Turk is less readily available) and those who do fieldwork working on less widely spoken languages. As a result, many researchers eschew formal experiments altogether, leading to a gulf between theoretical syntax methods and experimental syntax methods. Following Myers (2009)'s proposal of using mini-experiments to formalize and quantify the sort of informal syntactic exploration that happens anyway, we agree that there is space for a bridge between large-scale formal experiments and informal judgments. In order to address this issue, we propose the SNAP Judgments paradigm, which makes obtaining formal linguistic acceptability ratings easier and cheaper whether they are performed in the field, in the lab, or over the Internet.

5.2 Evaluation of Judgments from the Literature

We sampled a new set of sentences from the same 2001-2010 *Linguistic Inquiry* issues that SSA evaluated, and tested them in a rating experiment and a forced-choice experiment.

5.2.1 Experiment 1: Ratings Experiment

Participants

240 workers with US IP addresses were recruited through Amazon's Mechanical Turk crowdsourcing platform. 11 participants were excluded from the analysis because they did not self-identify as native speakers of English, leaving 229 participants. An additional 5 were excluded because they, on average, gave numerically higher ratings to the hypothesized unacceptable forms than to the hypothesized acceptable forms. These participants were thus probably not doing the task.

Stimuli

PG and an undergraduate assistant went through all the *Linguistic Inquiry* articles from 2001-2010 in which US English contrasts were presented and which were sampled from by SSA. We selected only contrasts that i) were English, ii) directly compared a grammatical and an ungrammatical sentence (irrespective of the particular judgment reported; assuming that OK>?>??>*?>*) and iii) were not primarily dependent on interpretation. This resulted in a total of 814 contrasts. Of these, 41 contrasts had already been tested by SSA. From the remaining 773 examples, we randomly sampled 101 contrasts.

For 96 of the contrasts sampled, Jeremy Hartman (JH; a syntactician) constructed a template illustrating which properties of sentences other than the syntactic parse were allowed to vary across experimental items and which were not. Next, JH constructed an example item based on the original contrast reported in the paper. 6 templates/sample item pairs were assigned to 16 MIT undergraduate students in MIT's 9.59J Laboratory in Psycholinguistics class (taught by Edward Gibson (EG); MIT 24.900 Introduction to Linguistics prerequisite). Students were asked to create 10 experimental items (hypothesized grammatical/ungrammatical pairs; 20 sentences total) for the contrast they were assigned. Student items were hand-checked and corrected by Peter Graff (PG) and JH. For each contrast, we tested the ten student items, the original sentence pair reported in the research article and JH's sample item, resulting in 12 experimental items per contrast. For the 5 additional contrasts in our sample, 11 items were constructed by JH, which, together with the original sentence pair reported in the paper, also resulted in 12 items for those 5 contrasts. All 1,212 contrast pairs were divided into 4 lists of 303 sentences each through a Latin square.

Procedure

Participants were asked to read each of the 606 sentences out loud to themselves and rate its naturalness on a Likert scale from 1-7. Order of presentation was randomized for every participant.

Results

After running the experiment, we noticed that 21 of our 1,212 sentences had minor spelling mistakes or errors in punctuation. We excluded these 21 sentences from the analysis reported below. These errors were fixed in the subsequent forced choice experiment. We also noticed that one contrast was constructed erroneously, in that it did not represent the intended contrast in its source article. We therefore excluded this item from the analysis, both here and in the forced choice experiment.

To eliminate some of the effect of different participants using the rating scale differently, ratings for each participant were z-transformed (mean and standard deviation estimated within participants). For each item in each contrast, we then calculated a mean z-score and averaged these together to get an overall z-score for the "acceptable" sentence and the "unacceptable" sentence in each contrast. The effect size is the difference between these two z-scores.

All 100 contrasts showed a numerical trend in the predicted direction. Following Sprouse & Almeida (2012), we computed Cohen's *d* for each contrast (Cohen 1994). Cohen's d is a measure of effect size that is equal to the difference in means between the two conditions (in this case, the acceptable condition vs. the unacceptable conditions) divided by the standard deviation of the data. Using Cohen's recommended effect size windows, we find 19/100 effects to be small effects ($d < .5$); 15/100 to be medium effects ($.5 < d < .8$); and 66/100 to be large effects ($d > .8$). Of the 19 small-effect contrasts, 7 actually have a Cohen's d of < .2, which is the minimum value that Cohen posits for a "small effect."

To control for individual variation by participant and item, we fit a linear mixed effects model with a sum-coded predictor for hypothesized acceptability (that is, whether or not the sentence is a reported as "acceptable" or "unacceptable" in the source *LI* paper) and random intercepts for both participant and item and random slopes for grammaticality grouped by both participant and item (random effects). The estimated coefficient for grammaticality is an estimate of the size of the effect, which is in this case the difference in z-score rating between the "acceptable" variant and the "unacceptable" variant, after controlling for participant and item effects. The model also gives us standard error estimates on this output. Figure 5-1 plots the effect size estimates and 95% confidence-intervals estimated from the mixed effect model. Despite the fact that many of the papers from which these examples were drawn talk about these contrasts categorically as either grammatical or ungrammatical, Figure 5-1 reveals that the effect sizes of randomly chosen linguistic judgments do not show any discrete jumps (which one might expect given the frequent discussions of "grammatical," "marginal," or "ungrammatical" sentences) but rather form a continuum from no effect to huge effects.

To assess significance in an NHST framework, we used the linear mixed effect model described above fit using the R statistical programming language (R Core Team, n.d.) and the lme4 package. We performed a χ_2-model likelihood ratio test comparing a model with a fixed effect for grammaticality to the intercept model (a model assuming a single mean for both hypothesized grammatical and ungrammatical sentences) leaving the random effects structure intact in both models (R. Baayen, 2008; Barr et al., 2013). This test asks whether the hypothesized grammaticality improves data likelihood significantly given the intercept, given normally distributed participant and item means, and normally distributed effect sizes of grammaticality for every participant and item. In other words, does hypothesized grammaticality explain a significant amount of variance in judgments?

To assess the statistical power of this experiment (the likelihood that we correctly

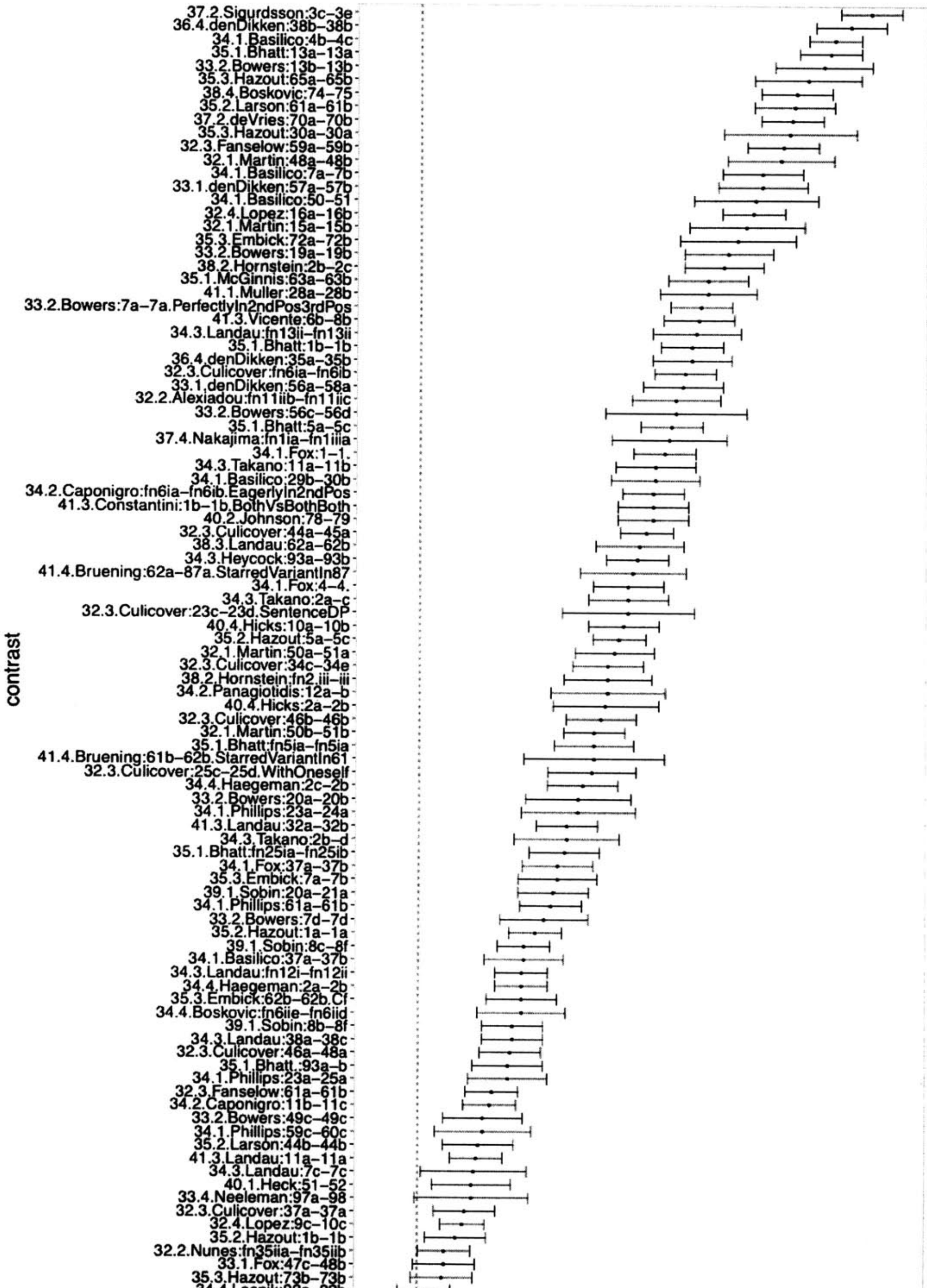


Figure 5-1: Each point is an effect size for the ratings experiment listed on the y-axis with 95% confidence intervals estimated from the linear mixed effect model. When the error bars extend through 0, the effect is not significant.

Table 5.1: Questionable results

Source article	Rating	Forced choice
35.3.Hazout.36-36	+	-*
34.4.Lasnik.24a-24b	+	-*
32.2.Nunes.fn35iia-fn35iib	+*	-
32.4.Lopez.9c-10c	+*	-
39.1.Sobin.8b-8f	+*	-
34.4.Lasnik.22a-22b	+	-
34.1.Basilico.11a-12a	+	+
35.3.Hazout:73b-73b	+	+*
33.1.Fox:47c-48b	+	+*
33.4.Neeleman:97a-98	+	+
34.3.Landau:7c-7c	+	+*

detect a true effect), we simulated results using the same number of subjects and items that were used in the analysis. The power analysis showed that, for a true effect size of 0.4, where effect size is the difference between the z-scores of the hypothesized grammatical and hypothesized ungrammatical sentences, we have a 96% chance of detecting a true effect at =.05. 81% of our contrasts are estimated to have effect sizes at least this big. For a true effect size of 0.2 (which would be small for experiments like these and would suggest very little difference between the two sentences), we have a 63% chance of detecting a true effect. The analysis shows that 92 of 100 contrasts in our random sample show significant effects in the predicted direction (92%). The contrasts that do not show clear effects are reported in Table 5.1 below.

Discussion

Of the 100 contrasts, the majority showed the predicted effect robustly. Of the eight that did not show a significant result, four showed clear trends in the predicted direction (35.3.Hazout:73b-73b, 33.1.Fox:47c-48b, 33.4.Neeleman:97a-98, 34.3.Landau:7c-7c). Four of the contrasts (35.3.Hazout:36-36, 34.4.Lasnik:24a-24b, 34.1.Basilico:11a-12a, 34.4.Lasnik:22a-22b) showed only numerical tendencies, with no clear trend in the predicted direction. We discuss the examples that did not show the predicted effects in more detail in Appendix D. It should not automatically be concluded that the inclusion of these sentences represent failures on the part of the researchers, although we do believe that these experiments suggest that these sentences warrant further investigation.

5.2.2 Experiment 2: Forced choice experiment

Participants

240 workers with US IP addresses were recruited through Amazon's Mechanical Turk crowdsourcing platform. We excluded participants who took the test more than once and those who did not self-identify as native speakers of English. We also excluded participants who chose the hypothesized "acceptable" option less than 60% of the time. Because most participants chose the "acceptable" option the vast majority of the time, those who were choosing it less than 60% of the time were likely not doing the task. After these exclusions, 201 participants remained.

Stimuli

The stimuli were the same as the stimuli in the ratings experiment, except with minor spelling corrections.

Procedure

Participants were asked to read each pair of sentences out loud to themselves and choose which sounded more natural. The order of presentation was randomized across participants.

Results

As in the rating experiment, one contrast was removed due to an error in how it was constructed, such that it did not represent the intended contrast in its source article. We found a wide array of effect sizes in the remaining sample of 100 contrasts, where effect size is taken to be the proportion of trials in which the hypothesized acceptable sentence is preferred. First, 6 of 100 contrasts trended in the opposite direction from the predicted direction. The remaining 94/100 sentences showed an effect in the predicted direction. 81/100 had an effect size greater than .75, and roughly half (52/100) had an effect size greater than .9. Overall, these results demonstrated smaller effects than those reported by SSA, but were qualitatively similar.

To control for individual variation by participant and item, we fit a logistic linear mixed effects model predicting whether or not the participant preferred the hypothesized acceptable sentence over the hypothesized unacceptable one. We included a fixed effect intercept (that is, whether or not the sentence was reported as acceptable or unacceptable in the original article) and random intercepts for both participant and item. The estimated coefficient for the intercept is essentially an estimate of how often a contrast would show a preference for the hypothesized acceptable form after controlling for participant and item effects. The model also gives standard errors for the estimates, from which we can calculate 95% confidence intervals (CIs). Figure 5-2 plots the effect size estimates and 95% CI output from the mixed effect model. As with the rating study, we see no discrete jumps but rather a continuum of effect sizes.

To assess significance in an NHST framework, we used the logistic linear mixed effect model described above and used the z-value to calculate a p-value. As before, we calculated statistical power for several possible true effect sizes. If the underlying effect size was 0.7 (meaning 70% of participants prefer the "good" sentence), for instance, we would have an 80% chance of detecting a true effect. The results appear in Appendix A at the end of this chapter. 92/100 showed significant trends in favor of the hypothesized acceptable version. Two items showed non-significant trends in favor of the hypothesized acceptable version, and four showed non-significant trends in favor of the hypothesized unacceptable version. Finally, two contrasts showed significant effects in the opposite direction of the predicted effect.

Discussion

In summary, 89/100 contrasts showed a significant effect in the predicted direction in both experiments, and 95/100 contrasts showed a significant effect in the predicted direction in at least one of the two experiments. These results therefore suggest that,

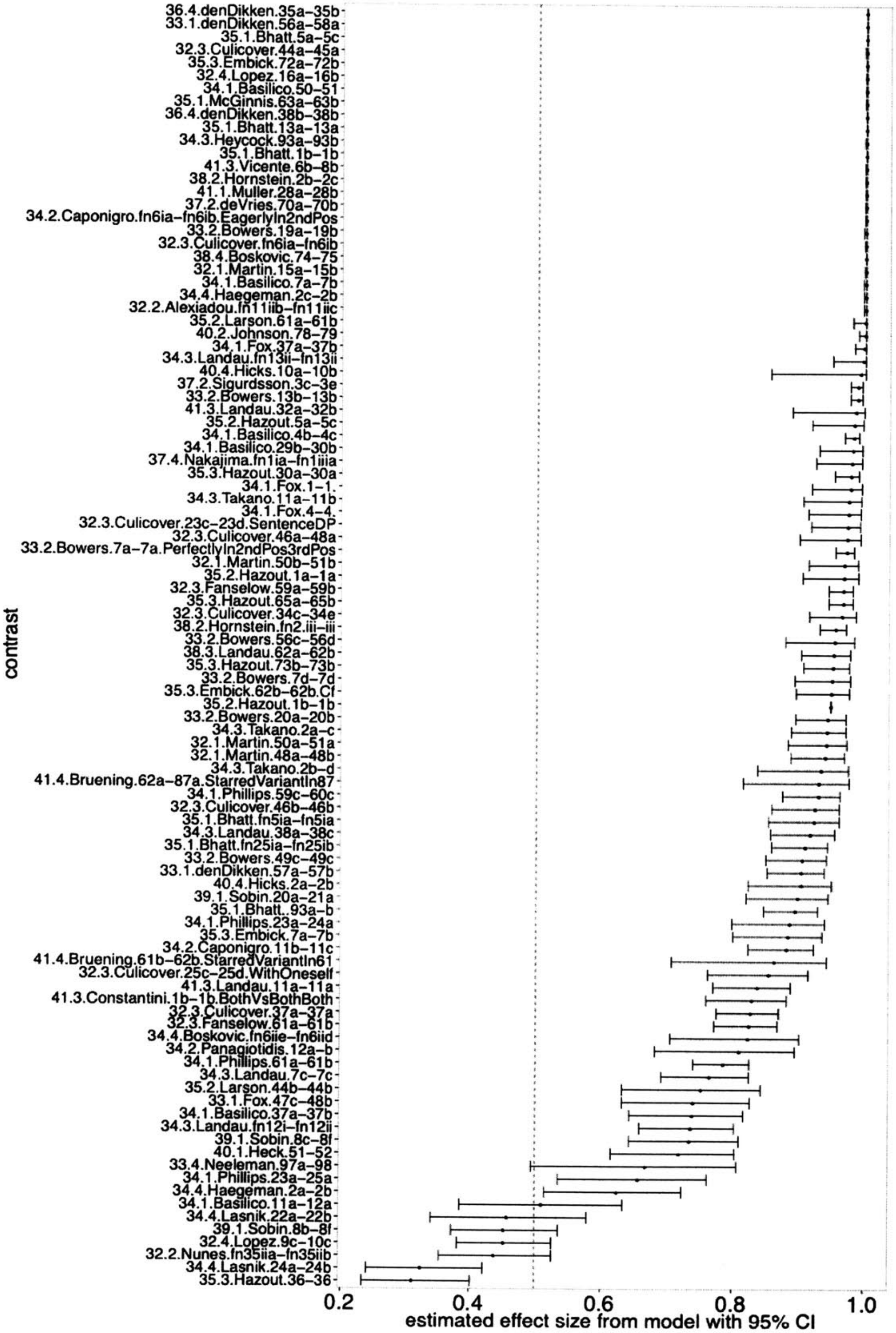


Figure 5-2: Each point is an effect size for the forced choice experiment listed on the y-axis with 95% confidence intervals estimated from the linear mixed effect model. When the error bars extend through 0, the effect is not significant.

whereas most published linguistic judgments are consistent with the results found by formal experiments, the methods are critically different.

Our results are quite similar to the results reported by SSA. The contrasts that did not show significant effects in the expected direction in our experiments are listed in Table 5.1, and we briefly discuss several of these cases in Appendix D.

5.3 SNAP Judgments

Across our two experimental paradigms, about half of the examples sampled from *LI* showed a strong effect (Cohen's d > .8). As many researchers have pointed out, it seems unnecessary to run a formal experiment in these cases, because they seem intuitively obvious. But it is not good scientific practice to rely only on intuition since even strong intuitions are sometimes wrong. To increase efficiency while maintaining scientific rigor, we propose a method for not having to run a large experiment while still being able to reach a statistically justified conclusion. We call this method a Small N Acceptability Paradigm for linguistic judgments (SNAP Judgments). We focus on forced choice judgments because they are simpler, have greater statistical power, and correspond more closely to the sorts of binary judgments between two sentences that often appear in linguistics papers. Of course, not all questions of acceptability are best investigated through forced choice judgments, and we leave it to future work to extend this method to other paradigms.

The basic idea is to be able to draw a statistically valid conclusion based on the data of just a few participants. To do so, we want to determine how many participants we would need to consult in order to be reasonably confident that we have a meaningful result. The simplest way to think of this is to think of each experimental trial as a flip of a weighted coin, where Heads corresponds to a preference for Sentence A and Tails corresponds to a preference for Sentence B. What we want to know is how often the coin will come up heads over a large sample of flips—that is, across many trials, how often Sentence A would be picked over Sentence B. We want to make inferences about the underlying parameter p, which is a probability between 0 and 1 that tells us how often Sentence A would be preferred over Sentence B. If p was .75, that means that 75% of the time, Sentence A would be preferred.

If we ask 5 people which sentence they prefer and they all prefer Sentence A, can we conclude with certainty that everyone will prefer Sentence A? No: it's possible that, if we ask a sixth person, she will prefer Sentence B. We still don't know if the probability of someone in the larger population preferring Sentence A is 90% or 70% or 10% (in which case getting 5 people who prefer Sentence A was an unlikely—but possible—accident). We want to try to infer p.

Here, we estimate p in a Bayesian framework by placing a beta prior distribution over the effect sizes found in our forced choice experiment above. In effect, this technique lets us supplement the result of our experiment by adding in prior information about what typical linguistic contrasts are like. If we were very confident that most effect sizes were large, then even after collecting just one data point, we might conclude that the effect was likely large. If we were very confident that the effect size was near .5, we might still think the effect is near .5 even if we asked 100 people and they all said the same thing. To decide how we should set our prior for linguistic judgments, we empirically estimated the parameters in the beta distribution by fitting the data we obtained in the Forced Choice experiment above. We found that the best prior was Beta(5.9, 1.1). This means that, for a new contrast where a researcher has used informal methods to decide that Sentence A is clearly better than Sentence B, this is roughly equivalent to an experiment where 6 participants have said they prefer A and 1 prefers B. (For a much more detailed description, see Appendix E.)

Using the prior, we can ask how many people we need to survey (while getting a unanimous result) in order to put our estimated effect size over .90 and our lower 95% confidence bound at or above .75. We chose this threshold somewhat arbitrarily based on our perception of what constitutes clear linguistic evidence. But this method can be used to ask what sample size is needed to achieve some other threshold by using the data in Table 5.2. Depending on the sort of conclusion a researcher plans to draw, different levels of effect size and confidence may be warranted.

If we ask 3 people whether they prefer Sentence A or Sentence B and they all say Sentence A, the expected mean proportion of people who would pick Sentence A is .89 with a 95% confidence interval of [.70, 1]. If we asked 5 people and they all answered Sentence A, the mean goes to .91, with a 95% CI [.75, 1]. Thus, where n=5 and the results of the experiment is unanimous, we get an expected mean of .91 and a lower 95% bound of .75. We believe that this is sufficient to be confident that the result would obtain in a larger experiment. Of course, as the sample size goes up, the experiment will give more information. Consequently, if a larger sample is easily obtainable, we recommend that. We do not recommend using fewer than 5 participants with 5 unique items for a SNAP Judgment.

5.3.1 Testing SNAP Judgments

We can test the efficacy of the SNAP Judgments proposal in the sample that we already have. First, we ask what the false positive rate is for SNAP Judgments: how often a SNAP Judgment will give a unanimous result when the underlying experiment is inconclusive. For each experiment, we randomly sampled 5 data points, each with a unique participant and a unique item. We then focused on only the experiments that gave us a unanimous result among the 5 randomly sampled data points. We can think of this as a simulated outcome for a SNAP Judgment. We repeated this procedure 100 times for each of the 100 experiments. On average, 54 of them produced a unanimous result. Of only those trials that produced a unanimous result, when we looked at the result of the full experiment with all participants and all items, the mean across those experiments was .92 with a 95% CI of [.76, .99]. Compare this to our Beta prior, which gave us an expected mean of .93 [.75, 1]. The empirical test is consistent with the results obtained using the Beta prior: the means match exactly, and the lower bound for the 95% CI is also very similar (.75 compared to .76 in the empirical test). Of the 54% of contrasts that pass SNAP Judgments, on average only 0.20% of them show results in the opposite direction of what is predicted and none of them included a significant result in the opposite direction.

Next, we ask about the statistical power of SNAP Judgments, specifically for large effects. That is, when an effect is very large, how often does the SNAP Judgment

Table 5.2: Means and CI's for sample sizes n=1 through n=10

n	**mean with 95% CI**	**mean with 99% CI**
1/1	0.87 [0.64, 1]	0.87 [0.5, 1]
2/2	0.88 [0.67, 1]	0.88 [0.55, 1]
3/3	0.89 [0.7, 1]	0.89 [0.59, 1]
4/4	0.9 [0.73, 1]	0.9 [0.62, 1]
5/5	0.91 [0.75, 1]	0.91 [0.65, 1]
6/6	0.92 [0.77, 1]	0.92 [0.67, 1]
7/7	0.92 [0.78, 1]	0.92 [0.69, 1]
8/8	0.93 [0.8, 1]	0.93 [0.71, 1]
9/9	0.93 [0.81, 1]	0.93 [0.73, 1]
10/10	0.94 [0.82, 1]	0.94 [0.74, 1]

paradigm fail to detect it and thus unnecessarily suggest a full experiment? We define a large effect here as one where the mean is greater than .90 (roughly half the experiments in our sample have this property). We simulate experiments as above, drawing on the real data from the subset of experiments where the mean is above the .90 threshold. On average, 77% of the experiments with true means above .90 produce unanimous effects in our simulations using 5 data points. Despite the small sample size, this is near the 80% threshold recommended for appropriate statistical power in experiments. Of course, for smaller effects, the statistical power will be much lower.

One may still, at this point, wonder why we recommend at least 5 data points for a SNAP Judgment as opposed to any arbitrary number. We believe that n=5, with a unanimous result, provides a robust generalization sufficient for most linguistic judgments. And crucially, it does not give a significant result in the wrong direction in any of the samples that we tested from *LI.* In Table 5.2, we give means and 95% and 99% CI's for SNAP Judgments that give unanimous results*in the predicted direction.*

5.3.2 Recommendation for SNAP Judgments

Given the proposal and evaluation above, we make the following recommendations for SNAP Judgments:

1. **To ensure the applicability of our empirical estimates, SNAP Judgments should only be used when the researcher believes, after informal investigation, that the effect is clear and likely to be unanimous.** If one does not believe that the results of the survey will be unanimous, it is better to do a large N rating study, which gives more gradient information, or a forced-choice study, which has more statistical power (Sprouse and Almeida 2012). From a statistical perspective, it is important that the researcher has this belief since the recommendations here are based on data that were published in a journal. Thus, if a researcher only has an inkling that Sentence A is better than Sentence B and wants to run a test to be sure, SNAP Judgments is not appropriate. In that case, we recommend a full, larger scale experiment.

2. **Construct 5 unique contrasts (each consisting of Sentence A vs. Sentence B, where one of the two sentences is hypothesized to be more acceptable than the other) and make sure that the 5 contrasts vary in lexical content and whatever other factors may influence the acceptability of the sentences in question. Present each contrast to a unique naïve participant and ask for a forced choice judgment.** This could be done using Amazon's Mechanical Turk (paying perhaps 5 cents for 1 judgment such that the whole experiment will cost less than 50 cents) or by simply asking for judgments from students, friends, informants, or colleagues who are naïve to the experiment in question.

Researchers in the field can use the same procedure, and in extreme situations in which access to speakers is severely limited (as in the case of endangered languages), it may be necessary to poll fewer than 5 participants. Table 5.2 below lists the conclusions that can be drawn from even smaller SNAP Judgments. For these extreme cases, we note that even the use of 3 independent data points substantially reduces the risk of a false positive. Thus, these recommendations need not dramatically slow or impede the pace of fieldwork: fieldworkers often do have 3 independent data points for a construction or contrast in question and can report those judgments quantitatively.

3. If all 5 naïve participants agree with the predicted result, one can conclude the following: the predicted mean for the full experiment is .91, with a 95% CI of [.75, 1] and a 99% CI of [.65, 1]. That is, one can be 95% confident that at least 75% of people would agree with the intuition. See Table 5.2 for guidelines when using participant sample sizes other than 5.

4. If the judgments are not all in agreement, then the intuition that the result would be unanimous is wrong. This is not a failure of SNAP Judgments but one of its major advantages: giving the opportunity to explore where there is variation among items. If there is not agreement among participants, look at the five items. Is there a pattern among items that do not show the expected effect? Do variations in word choice or prosody or context seem to affect the results? If so, this might be an opportunity to further explore the nature and size of the effect in question. If new hypotheses are generated by the SNAP Judgment experiment, one can then test these hypotheses. At that point, we recommend a formal experiment with a larger number of items and participants to understand the size of the effect and possible sources of variation. (Note that, if one were to simply perform SNAP Judgments repeatedly on the same grammatical contrast, one might eventually find a string of unanimous responses just by chance. This is the multiple comparison fallacy, and, in that case, the statistical guidelines here would not be directly applicable.)

5.3.3 Limitations of SNAP Judgments

Although the SNAP Judgments paradigm addresses issues with statistical power in running linguistic experiments, there are a number of limitations of this technique that are worth considering and that can be explored in future research. For one, the guidelines presented here do not solve the problem of how to write good items that generalize to the contrast in question. Even if a researcher were to test 100 versions of some specific syntactic generalization, she may have overlooked some special case in which the generalization does not hold due to lexical, pragmatic, or contextual factors. Statistics should supplement, not replace, careful thought about syntax and semantics.

Moreover, the SNAP Judgment proposal does not address what factors are involved in distinguishing the acceptability of two structures. In particular, this framework does not guarantee that any differences are due to syntax ("grammaticality") or to some other factor that might systematically differ between the two structures, such as world knowledge ("plausibility") or lexical or discourse properties.

When SNAP Judgments is used with in-person participants as opposed to over the Internet, we recommend caution in making sure that the researcher not bias the participants towards any particular answer. When done online using a crowdsourcing service like Mechanical Turk, it is good practice to include some checks to make sure that participants are engaged, doing the task, and competent users of the language of interest. For more on using Mechanical Turk for language research, see Crump, McDonnell, and Gureckis (2013); Gibson, Piantadosi, and Fedorenko (2011); Mason and Suri (2012); Sprouse (2011).

Another limitation of the paradigm here is that we have so far tested it only for English sentences sampled from *Linguistic Inquiry*. More work is needed to see how well the recommendations here extend to other languages and other types of sentences that may be of theoretical interest. There is also more work needed to know whether *LI* sentences are typical of judgments of theoretical interest or whether they differ in meaningful ways from other peer-reviewed journals and whether journals differ from conference proceedings.

Finally, it is important to note that the judgments we trained on here are published judgments and thus not necessarily representative of the judgments that linguists are faced with in day-to-day research. In particular, because the recommendations here are based on published judgments that the authors presumably believe are correct, the model assumptions are not valid for questions where a researcher is uncertain. It is possible that the distribution observed here would be different using unpublished data and thus that there might be different SNAP Judgment recommendations depending on where in the publication pipeline a particular judgment is.

5.4 Discussion

In this paper, we have replicated the empirical findings of SSA. In a sample of 100 contrasts from *Linguistic Inquiry,* we found a wide range of effect sizes in both a rating experiment and a forced choice experiment. Small, medium, and large effects are all well represented in the data set. 89% of these syntactic judgments reported in *LI* show significant effects in the predicted direction in two types of large-scale formal experiments. In the remaining 11%, there are varying levels of uncertainty about the judgments elicited. In all of these cases, we believe that the formal experiments uncover interesting sources of variation that could illuminate the theoretical questions at stake and improve the papers in which they appeared.

Moreover, we used the empirical results presented here as a foundation on which to build a prior distribution of what syntactic judgments can be expected to look like. Specifically, we found that a Beta distribution is a good fit to the distribution of probabilities found in the forced-choice experiment and used it to recommend a new paradigm for small-sample acceptability experiments.

We believe that the SNAP Judgments paradigm will make it easier and cheaper for language researchers to obtain statistically justified linguistic acceptability judgments. Specifically, in instances where a researcher is confident that a judgment would produce a unanimous result across 5 participants, we recommend a forced-choice experiment with 5 participants and 5 items. If the result is unanimous, the results of this small N experiment can be combined with a beta prior to give a predicted effect size of .93 with a 95% CI [.75, 1].

There is great value in being able to attain cheap and easy quantitative data in syntax and semantics–and in knowing when to run larger experiments. Marantz (2005) (and more recently Linzen and Oseki 2015) proposes three distinct classes of acceptability judgment: judgments that contrast word salad with obviously grammatical language ("Ate rat cat the." vs. "The cat ate the rat."), judgments that test "obvious" features of a grammar such as adding –*ed* to form a past tense in English, and judgments that explore more subtle features of a language like reference or long-distance dependency. Broadly, this three-way categorization classifies judgments into "obvious" (Marantz's first two classes) and "non-obvious" judgments (the third class). Marantz suggests that this third class of more subtle judgments (what we call "non-obvious") are the ones that could benefit from more formal experimentation. Linzen and Oseki (2015) show that these intuitively non-obvious judgments are much less likely to be replicated in a formal experiment, with failures to replicate of 50% in Hebrew and 32% in Japanese for these non-obvious judgments (where obviousness was coded by the experimenters).

But how does a researcher know *a priori* if she is dealing with a non-obvious judgment? 5 independent coders (K.M., E.G., and 3 others) annotated the 100 *LI* contrasts from our sample and classified them as obvious or non-obvious and found that the average agreement between two coders was only 68%. Thus, it is not always obvious when a contrast is obvious. The authors discussed the disagreements among the five raters and arrived at a consensus set of obvious and non-obvious judgments, without consulting the experimental data for these contrasts. Here, the contrasts that were rated as "obvious" had on average a mean of 92% in the forced choice experiments, compared to 79% for the non-obvious judgments. 98% of "obvious" judgments showed significant results in the predicted direction in the forced choice full experiment, compared to just 86% for the "non-obvious" judgments. Because judgments with over 90% agreement in a full experiment were shown to be likely to pass SNAP Judgments, the SNAP Judgments paradigm gives researchers an effective way to know if they are dealing with a judgment that is in need of formal experimentation: judgments that do not produce unanimous SNAP results are likely not "obvious" contrasts. To that end, one way of thinking of SNAP Judgments is as an empirical procedure for determining whether a syntactic contrast falls into the non-obvious class of judgments that warrant further formal experimentation.

SNAP Judgments will lead to more published quantitative data in two ways: (a) through the publication of small-sample SNAP Judgment data and (b) through the increased use of formal acceptability experiments for contrasts that do not meet the SNAP threshold. Not only will the collection of more quantitative data help preclude erroneous analyses from entering the literature, but it will also enable us to continue building a body of empirical data on acceptability judgments. This body of data will make it easy for researchers to uncover new and interesting empirical phenomena and place those phenomena into a larger quantitative framework so as to understand gradient effects and sources of variation in linguistic data. For instance, using the data from this study (available for download on the Open Science Foundation at the URL osf.io/5wm2a), one can easily test a new contrast and plot it alongside the 100 phenomena tested here in order to ask what other sentences the contrast patterns like, how much variation there is among participants for that contrast, and how sensitive the contrast is to variation in lexical items or context. Knowing whether a particular proposed effect is small, medium or large–and knowing exactly what that means relative to other published judgments—is a worthwhile goal.

Given the ease with which SNAP Judgments can be attained, we believe that the time and effort required is not much more than what a researcher already spends when discussing judgments with friends, colleagues, and students or what a field linguist spends eliciting judgments from informants. By treating syntax and semantics questions empirically, we can develop standardized quantitative methods that can be shared across disciplines, across languages, and by future generations of researchers.

5.5 Appendix A: Rating Study Results

See full results at Open Science Foundation URL osf.io/5wm2a.

5.6 Appendix B: Forced Choice Results

See full results at Open Science Foundation URL osf.io/5wm2a.

5.7 Appendix C. References

See full set of materials in the Materials folder at the Open Science Foundation URL osf.io/5wm2a.

5.8 Appendix D: Discussion of items that do not show clear results in the predicted direction.

35.3 Hazout.36:

(#) There seem/*seems to have appeared [some new candidates] in the course of the presidential campaign.

The rating study revealed no significant difference between the two variants (=0), and the starred variant was significantly preferred in the forced choice experiment. This judgment seems to reflect a trend in colloquial English to use the singular "There seems" in these "verbal existential sentences," even when the agreeing phrase is plural. At the very least, there may be individual variation in sentences like this.

34.4.Lasnik.24a-24b:

a. ?The detective asserted two students to have been at the demonstration during each other's hearings.

b. ?*The detective asserted that two students were at the demonstration during each other's hearings.

(b) is proposed to be unacceptable only when the final PP modifies the matrix clause and not the embedded clause. Our items were written to ensure that this is the only plausible interpretation, but participants still preferred (b) by a significant margin in the forced-choice experiment.

34.4.Lasnik.22a-22b:

a. John proved three chapters to have been plagiarized with one convincing example each.

b. ?*John proved that three chapters were plagiarized with one convincing example each.

This example showed a non-significant trend in favor of (a) in the rating study and a non-significant trend towards (b) in the forced choice study. Again, we took care to ensure that the final PP modifies the matrix verb across all our items.

32.4.Lopez.9c-10c

a. We proved Smith to the authorities to be the thief

b. *We proved to the authorities Smith to be the thief.

People significantly preferred (a) in the rating study, but the opposite trend emerged in the forced choice study, which suggests that this is not a clear contrast. In fact, Hartman (2011) has argued that sentences like (a) are degraded on independent grounds, which might explain why most subjects did not prefer them over (b).

39.1.Sobin.8b-8f

a. Bill devoured a ham, and Mary did a similar thing with a chicken.

b. *Bill devoured a ham, and Mary did so with a chicken.

In this contrast, we found a significant predicted effect in the rating study but a trend in the opposite direction in the forced choice experiment. It is possible, in this case, that the "did so" construction in (b) is semantically unclear out of context, but clearer (and more natural sounding) when presented with the more semantically transparent (a). This would explain the difference between the rating study and the forced choice study.

5.9 Appendix E: Math behind SNAP Judgments

Formally, we can think of our experiment as a draw from a binomial distribution, where p is the underlying population parameter for how likely someone is to choose Sentence A over Sentence B, n is the total number of trials, and k is the number of trials on which someone chose Sentence A over Sentence B.

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{5.1}$$

To obtain a confidence interval from a binomial distribution where the sample is unanimous while also taking advantage of our prior knowledge about how *most* experiments turn out, we will use a Bayesian credible interval—which is the Bayesian version of a confidence interval and can be thought of as the probability that a given parameter falls within some interval—on the posterior distribution. We get the posterior distribution by combining our binomial likelihood with a Beta prior distribution (Gelman et al. 2004) on the parameter p, which gives a distribution of possible values for our parameter p. This prior distribution is the distribution over the value of p is *before* we have collected any data. In other words, before we flip the coin, we do not know its weight p. We might think that it is very likely that the coin is fair and that p is near .50. Or maybe we think that p is close to 1. The shape of the distribution is controlled by the shape parameters and . Formally, the beta distribution is:

$$P(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\beta(\alpha,\beta)} \tag{5.2}$$

where B is the beta function. We could, in principle use any distribution with support on [0,1], but we use the beta distribution because it is the conjugate prior for the binomial and thus lets us obtain a closed form solution.

Informally, we can think of the job of the prior as being to add in our prior belief about the underlying distribution. We can literally think of this as adding the results of imaginary trials that we have not actually conducted. For instance, if we suspect that the coin is fair, we might use a Beta prior of Beta(5,5)—meaning and are both 5. Then, we present 5 people with Sentence A and Sentence B and ask which is better. In this case, p is the underlying probability of choosing A. We get the following result:

A A A A A

Without the prior, our best guess for the underlying parameter p is 1 since 5/5 is 1. If we use the Beta(5, 5) prior, however, we can think of this as adding 5 *a priori* A's and 5 *a priori* B's to our 5 experimentally obtained A's such that we imagine we have 10 A's and 5 B's, as in the following (where the italicized values come from the prior):

A A A A A B B B B B A A A A A

In this case, our best estimate of the underlying parameter p is (5 A's + 5 A's) /

(15 trials) = 66. If we were very confident that the sentences are equally acceptable (i.e., the coin is fair; p ~ .5), we could use a Beta(100, 100) prior. With a prior like that, we'd have to conduct many more trials in order to move our estimate substantially away from .50. After getting 5 A's, we would still have an estimate of 51%.

If we thought it was very likely that one of the sentences was better, but we didn't know which, we might instead use a Beta prior of Beta(.1,.1). This would mean that, after asking 5 people who all choose A, our new estimate for how likely a random person is to choose A would be: 5.1/(5.1 + .1) = 98%. Figure 3 shows the shape of the beta distribution for 2 possible settings of the shape parameters. If the shape parameters are unequal, then the distribution is skewed. When the two shape parameters are equal, the distribution is symmetric.

Formally, we can multiply the beta prior and the binomial likelihood together to get the posterior probability.

$$P(k|n,p) * P(p|\alpha,\beta) = P(k|n,\alpha,\beta) = \binom{n}{k}\frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)} \qquad (5.3)$$

In our case, we want to know what our prior expectations about p should be. Should our prior look more like Figure 5-3.a or Figure 5-3.b? Because we have formal results for 100 contrasts, we can use these empirical results to set our prior. In other words, when we have a new contrast for which we don't have much data but which we believe likely to produce a unanimous result, we can imagine that the contrast has an underlying parameter p (where p is once again the probability of choosing Sentence A) and that p is drawn from the same distribution of judgments that gave rise to the 100 contrasts we observed. If we do not believe that the contrast is likely to produce a unanimous result,

the assumption that the parameter p is drawn from the same distribution as the 100 contrasts we tested experimentally is potentially invalid since, in general, the effects that we tested were hypothesized to be very strong.

In order to determine the prior empirically, for each contrast in our experiment, we assume that the hypothesized "good" sentence is Sentence A. We then draw a

Figure 5-3: The histograms represent a density map of a draw from a beta distribution with the shape parameters indicated. The red line is the probability density of the beta distribution at each value for p between 0 and 1. The plot on th left conforms to an instance in which, most of the time, the probability p is extreme (towards 0 or 1), as in the experiments we tested here. The plot on the right corresponds to a situation in which we have a strong prior belief that the probability p is near .5.

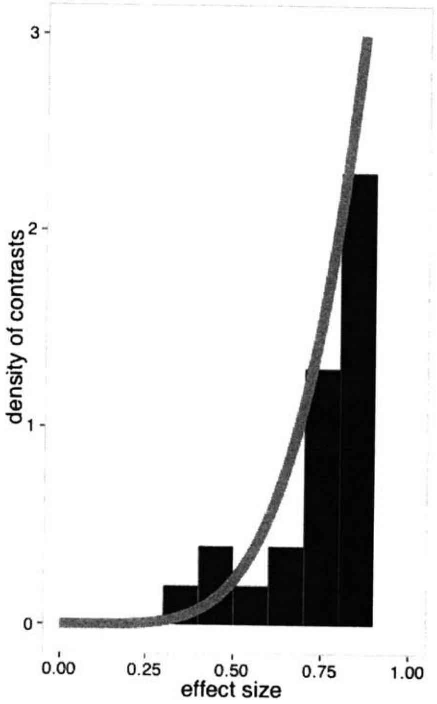


Figure 5-4: This plot corresponds to a smoothed histogram (averaged over many trials) of the data from our forced choice experiment where, for each contrast, one variant is randomly assigned to be Sentence A and one to be Sentence B. Most of the time, there is a strong preference for one sentence or the other. The best fit for the beta distribution is Beta(5.9, 1.1)—which is shown by the red line.

histogram of the effect sizes and fit the beta distribution to the histogram (as seen in Figure 5-4). Averaging over 100 samples, the best fit is Beta(5.9, 1.1) with standard error .12 and .01 respectively. Rounding to the nearest whole number, we can think of this as having seen 6 A's and 1 B *before* we run our experiment. Thus, if we run an experiment and get 3 A's and 0 B's, we can act as if we have 9 A's and 1 B. We can use this prior to construct 95% Bayesian credible intervals for the underlying probability in the population of someone preferring Sentence A over Sentence B. Specifically, the Bayesian credible interval gives us a continuous interval, for which there is a 95% probability that the true underlying probability falls in that region.

We also checked to see if the recommendations here were robust to other reasonable choices of prior. There is some theoretical question as to whether it makes sense to use the full available information in order to set the prior or if we should instead "forget" which sentence is hypothesized to be good and assume that it is equally likely that the good sentence is A or B. The logic here is that including information as to which sentence is supposed to be good would be equivalent to doing an experiment where a researcher wants to test the efficacy of a medicine and then includes her prior belief that the medicine will probably work as evidence in the experiment. While she might be very confident in the medicine's efficacy, she cannot include that prior belief as part of her analysis or else she could end up concluding that data which are consistent with pure noise is actually a result in favor of the hypothesis. But, because the whole point of the SNAP Judgment paradigm is to use the existing information, we do not believe those concerns are particularly relevant here.

To check how robust the paradigm is to choice of prior, we tried this approach where A and B are equally likely to be the "good" sentence. To do that, we randomly assign one sentence in each contrast as A and one as B. Using this approach, we find a Beta(.6, .6) prior. For 5 unanimous participants, this gives us a mean of .90 with a 95% CI of [.67, 1]. So the CI's lower bound is only slightly lower than when we include all the information. To get the lower bound to .75 when we use this prior, we would need to include 7 participants in the experiment (as compared to 5 in our main analysis). We would also arrive at similar conclusions if we used the Jeffreys uninformative prior Beta(.5, .5)—a prior that is standardly used in many applications since it locally uniform. Hence, the outcome is similar under other plausible alternative priors. We use the asymmetric, full-information prior in our main analysis, but we recognize that there may be good theoretical reasons to instead use the symmetric prior.

5.10 Appendix F. Statistical Power

The idea of computing statistical power is to ask, if there is an underlying "true effect" size D that is being looked for in the experiment, what is the likelihood that the experiment correctly detects a significant effect? (Note that, in reality, we can never know the "true effect size" because that would require infinite data. We can only sample.) If $D = .8$ for a sentence in the forced choice experiment, that would mean the true underlying effect was .80. If the statistical power of our experiment .95 (based on the sample size and design), that would mean that 95% of the time we would find a significant effect given the underlying effect size of .80. (Power would be lower if the effect size were smaller.) To compute statistical power and possible error rates using linear mixed effect models, we repeated the following procedure 100 times for each contrast, took the mean of those 100 iterations, and then averaged across contrasts.

a) Fit a linear mixed effect model to the real data as described in the main text.

b) Use the random effect structure and residual variance from the model fit to the actual data in a). For the fixed effect estimate, use D which we systematically vary and report for several values in the table below. In effect, this lets us use the actual variance in the world (by subject, by item, and residual variance) to estimate the noise we should expect in an experiment.

c) Use the parameters from b) to simulate a new set of data equivalent in sample size to the original experiment and with the same subject and item breakdown as the original experiment.

d) Fit a new linear mixed effect model to the simulated data in c) and test for effect size and significance.

Table 5.3: Power and error rates for ratings studies

D (true effect size)	Statistical power	Type S error rate	Type M error rate
.2	0.63	0.0	1.29
.4	0.96	0.0	1.01
.6	1.00	0.0	1.00

e) Use the effect sizes and significance levels found in d) to calculate power, Type S, and Type M error.

We used the simulated effect size and significance measures to calculate statistical power given varying underlying effect sizes as well as two measures recommended: Type S (Sign) Error and Type M (Magnitude) Error (Gelman and Carlin 2014). Power here refers to the proportion of the time a "true effect" would be detected in the experiment given true effect size *D*. Type S error refers to the proportion of the time a significant effect is found in the *opposite* direction of the true effect. That is, if the Type S error rate is .05, that means that 5% of the time, we should expect to find a significant effect in the opposite direction of the true effect. Type M error refers to the expected absolute overestimation rate given that a significant effect is found (that is, when significant, the absolute value of the estimated effect size divided by the true effect size). This means that, conditioned on finding a significant effect, we should expect it to be *M* times more extreme than the underlying true effect.

The below tables report power and estimated error rates for various true effect sizes. Note that, in the rating study, a true effect size less than .4 is quite small (only 19% of our estimated effect sizes are this small) and possibly not large enough for robust acceptability generalizations. For the forced choice study, an effect size less than .70 is quite small and only 11% of our data fits that description.

Table 5.4: Power and error rates for forced choice studies

D (true effect size)	**Statistical power**	**Type S error rate**	**Type M error rate**
.6	.48	.04	1.71
.7	.80	0.0	1.17
.8	.93	0.0	1.06

Chapter 6

Evaluating methods in psycholinguistics: Meta-analysis of syntactic priming

6.1 Introduction

The previous chapter dealt with issues in linguistic acceptability judgments and advocated for a more data-driven, experimental approach. Here, we evaluate how an experimental approach has done at building up knowledge by focusing on the last 30 years of studies on syntactic priming.

When someone is primed with some syntactic structure X and is then asked to produce a new sentence, it is claimed that she is more likely to use that same structure X than if she had instead heard some other structure Y. This phenomenon, *syntactic priming* (also sometimes called *structural priming* or *syntactic persistence*), has been an important topic of study in psycholinguistics since Bock (1986). Its existence has been used as evidence for the abstract representation of syntactic structures in the mind, and it has gone on to be used in hundreds of experiments. See Ferreira & Bock (2006), Heydel & Murray (2000), and Pickering & Ferreira (2008) for critical reviews of this literature.

As a phenomenon that has become quite central to the field of psycholinguistics, with hundreds of papers published on it in the last 30 years, syntactic priming is ripe for a cumulative quantitative analysis. Indeed, syntactic priming has been used to test theories of event structure (Bunger, Papafragou, & Trueswell, 2013), social interaction (Branigan, Pickering, McLean, & Cleland, 2007), bilingualism (Bernolet, Hartsuiker, & Pickering, 2007, 2013); Schoonbaert, Hartsuiker, & Pickering, (Schoonbaert, Hartsuiker, & Pickering, 2007), syntactic surprisal (Jaeger & Snider, (T. F. Jaeger & Snider, 2013), childhood linguistic representations (Messenger, (Messenger, 2010), amnesia (Ferreira, Bock, Wilson, & Cohen, (V. Ferreira, Bock, Wilson, & Cohen, 2008), autism (Slocombe et al., (Slocombe et al., 2013), aphasia (Verreyt et al., (Verreyt et al., 2013), implicit learning (Kaschak, Kutta, & Jones, (Kaschak, Kutta, & Jones, 2011), and human mating behavior (Coyle & Kaschak, (Coyle & Kaschak, 2012). Perhaps most critically, syntactic priming has been used as evidence for the abstractness of syntactic operations (Bock, 1986, 1989)–one of the core claims of modern linguistics. Thus, one of the goals of this meta-analysis is to assess and help organize the current state of knowledge in the field by aggregating data and evaluating it quantitatively. All else equal, how big is the syntactic priming effect? What is the range of variation one could expect? How much bigger should it be when there is lexical overlap between the prime and target? Can the existing literature be trusted, or does it likely suffer from publication bias?

While there have been several large-scale critical reviews of syntactic priming, there has not been a systematic, large-scale quantitative meta-analysis (but see Jaeger & Snider (2013) for a meta-analysis of three earlier experiments). Meta-analyses, whereby a group of studies are gathered and quantitatively analyzed together, can be useful for assessing what we have learned through a large body of distinct studies and for exploring whether these studies are exploring the same underlying phenomena citemark2001practical. Meta-analyses dramatically increase statistical power–that is, the probability of detecting a true effect–by pooling data together. In our meta-analysis, for instance, we included data from over 5,000 unique participants, whereas no single experiment in our sample used more than 144. For these reasons, increased use of meta-analysis in the social sciences has been widely recommended as a way to investigate the reliability of published results (Button et al., 2013b, 2013b; Cumming, 2013; Simonsohn et al., 2014a, 2014a).

In this paper, we report three results: a standard meta-analysis, an analysis of publication bias, and recommendations for sample size in future priming studies. We take effect size to be the log odds ratio of the proportion of target structure produced in the prime condition to the proportion of target structure produced in the no-prime condition. For 45 of the 73 papers in our sample, we obtained raw data from the authors and used it to derive estimates of effect size and standard error. From the remaining papers, we estimated the effect size and standard error using the published estimates. Along with effect sizes and their associated standard errors, we also collected information on several key manipulations that can potentially modulate the priming effect, including the construction used, lexical repetition, lag, and whether the priming is within or across languages. Using these variables, we estimated the average effect size of syntactic priming given various experimental conditions.

As a secondary analysis, we assessed the extent to which the set of papers in our study suffer from publication bias and low power. Indeed, there have been meta-analyses in other branches of psychology alleging widespread publication bias (Ioannidis et al., 2014; Landy & Goodwin, 2014), low reproducibility (Open Science Collaboration, 2015), and low statistical power (Button et al., 2013b, 2013a). Low statistical power can lead to inflated false-positive rates in the literature and unreliable results (Gelman & Carlin, 2014). To assess publication bias and statistical power, we used p-curve, a tool developed for that purpose which works by analyzing the distribution of significant p-values in the literature (Simonsohn et al., 2014a, 2014b). Using the raw data gathered from the study authors, we did a power analysis and give guidelines on how to run syntactic priming studies with sufficient statistical power.

Besides just investigating and quantifying the state of the field, though, there are a number of open questions in syntactic priming that we can investigate using this method. For instance, Pickering & Ferreira (2008) describe conflicting evidence as to just how long lived syntactic priming is. Here, we provide evidence that, as Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst (2008) suggests, syntactic priming decays relatively slowly but the effect of lexical overlap decays quickly. We also show that, when there is lexical overlap between the prime and target, syntactic priming is very strong in a speaker's second language–much stronger than any observed priming within a first language. This collection of priming results, analyzed together for the first time, yields the strongest support yet to claims that priming is an abstract process largely independent of modality or task.

6.2 Meta-analysis

6.2.1 Method

For our main meta-analysis, we exhaustively searched for a set of papers on syntactic priming in production. We then extracted the measures of effect size along with details of the experimental set-up. Finally, we performed several regressions to assess (a) the size of the overall priming effect and (b) how it is affected by variations in the experimental conditions.

Inclusion and exclusion criteria

We included only controlled experiments that were focused on syntactic priming in production, in which prime sentences were designed to elicit participant-generated productions of the same syntactic structure and in which the dependent variable was the production itself (thus excluding studies where the dependent variable is reaction time or some other psychometric measure). We defined "syntactic priming" as priming above the level of the word and as not including priming of inflectional or derivational morphology or metrical structure. For that reason, we will refer to the phenomenon mostly as "syntactic priming" (as opposed to "structural priming" or "structural persistence"). While there are many arguments that could be made as to what constitutes a *syntactic* alternation as opposed to a lexical, semantic or pragmatic one, we enforce a fairly strict requirement that the alternation involve a change in word order that does not dramatically alter the semantics of the utterance.

Classic alternations are the active/passive alternation ("The boy chased the ball." vs. "The ball was chased by the boy.") and the dative alternation ("The man gave the boy the ball." vs. "The man gave the ball to the boy."). We excluded experiments like Verreyt et al. (2013) that involved the priming of thematic roles since this did not strictly prime word order. While studies like these likely tap into syntax, we chose to be narrow in our inclusions in order to have a more homogenous sample.

We further constrained our sample to experiments with healthy adult participants. When a study focuses on a non-healthy or child population but also presents data from a control group of healthy adult participants, we included the control group in our sample. We required that the results be published in English-language, peer-reviewed journals (not including conference proceedings or dissertations) in 2013 or earlier. Finally, we only considered papers where the proportion of productions matching the primed structures was included as a dependent measure.

Our criteria exclude some studies which are sometimes classified as syntactic priming. Specifically, we chose not to include comprehension priming studies, which includes any study where the dependent measure is not a linguistic production but a measure of how a priming manipulation affects participants' comprehension of sentences (but see Tooley & Traxler (2010) for a review of the extensive literature on comprehension priming). We excluded recall studies (e.g., Potter & Lombardi (1998) in which a participant is asked to recall a memorized sentence and the dependent measure is whether she makes errors) since these studies differ from other priming production studies in that they have a "right" answer and do not encourage the same sort of free production. For the same reason, we excluded cross-linguistic priming studies in which the task is to directly translate a sentence from one language to the other. We excluded studies that were not strictly controlled–including "syntactic alignment" studies whereby the dependent measure is how well the use of a particular structure X predicts the use of that structure at a later time in free-form conversation. While alignment of this nature is arguably a subset of priming, it is beyond the scope of this meta-analysis.

These criteria were applied to 2,096 records returned during the search process, resulting in 73 records.

Moderator analysis

In addition to the main priming effect, studies often investigate other questions about the mechanisms underlying structural priming. Because it is not possible to model all possible variation between experiments, meta-analysis requires choosing a number of experimental variables to consider as *moderators* of the priming effect. For instance, a body of literature has investigated whether there are medium-to-long term effects of structural priming. There are a variety of such moderator variables, and we will include some of these variables as predictors in the meta-analysis.

We extracted information from each paper for the moderators listed below. Each bullet point is a particular variable, and each sub-bullet point represents possible values for that variable. For some variables, we recorded more detail but collapsed them into the possible values shown below.

- Language
- Construction type
  - active/passive: "The boy kicked the ball." vs. "The boy was kicked by the ball." See, e.g., Bock (1986).
  - complex NP: "The man in the car..." vs. "The man...." See, e.g., Bunger et al. (2013).
  - dative: "The girl gave the boy a ball." vs. "The girl gave a ball to the boy." See, e.g., Bock (1986).
  - genitive: "The man's car" vs. "The car of the man." See, e.g., Bernolet, Hartsuiker, & Pickering (2012).
  - transitive/intransitive: "The man was driving." vs. "The man was driving the car." See, e.g., R. P. van Gompel, Arai, & Pearson (2012).
  - locative: "A cat lies on the table." vs. "On the table lies a cat." See, e.g., Hartsuiker (1999).

  - modifer order (preferred vs. dispreferred): "The big, red chair" vs. "The red, big chair." See, e.g., Goudbeek & Krahmer (2012).
  - NP (complex or simple modifier): "The red book" vs. "The book that's red." See, e.g., Cleland (2003).
  - R.C. attachment (high or low): Relative clause attachment, as in "The men with the kids who plays the piano" vs. "The men with the kids who play the piano." See, e.g., Scheepers (2003).
  - verb-participle order: "De man belde de politie omdat zijn portemonnee was gestolen/The man called the police, because his wallet was stolen." vs. "De man belde de politie omdat zijn portemonnee gestolen was/The man called the police, because his wallet stolen was." (from Hartsuiker & Westenberg (2000)).
  - VP syntax: "The woman entered a cave." vs. "The woman drove into a cave." See, e.g., Bunger et al. (2013).

- Temporal lag between the prime and the target.
  - none: Target appears after prime with no intervening linguistic material (there can be a fixation screen).
  - filler: Some number of filler items appear between the prime and target.
  - cumulative block priming: In this category, we included studies in which the prime does not precede the target in an alternating fashion, but rather the priming round occurs followed by a target round (i.e. all the primes occur, then all the targets). Many of the Kaschak, et al. studies (e.g., Kaschak et al. (2011), Kaschak, Loney, & Borreggine (2006)) fall into this category.
  - cumulative block priming (long): A priming round occurs more than 10 minutes before a target round (e.g., Kaschak (2007)). We chose to separate this category from other types of cumulative priming since we

hypothesized that long delays (including several days in some studies) could have a qualitatively different effect than priming on the scale of seconds or minutes.

- Bilingualism
  - L1 → L1 (priming within first language)
  - L2 → L2 (priming within second language)
  - L1 → L2 priming (cross-linguistic with first-language prime, second-language target)
  - L2 → L1 priming (cross-linguistic with second-language prime, first-language traget)
- Lexical overlap between the prime and target
  - No if no repeat of critical words between prime and target.
  - Yes otherwise (if word is repeated, if semantically related word is repeated, if translated version of word is repeated, etc.).
- Year of publication
- Target task
  - Picture description (participant orally describes a picture).
  - Written sentence completion (participant fills in a sentence like "The boy gave...").
  - Auditory sentence completion (participant speaks the completion to a sentence like "The boy gave...").
  - Sentence from words (participant is given target words and told to assemble them into a sentence).
- Modality of prime

  - Auditory prime (including both recordings as well as spoken primes by a "live" interlocutor).
  - Visually presented prime.
  - Prime is visually presented but read aloud by participant.
- Whether the prime is repeated by the participant
  - Yes (includes cases where the prime is delivered auditorily and then repeated as well as cases where the participant self-primes by being required to complete a prime sentence in a particular way)
  - No
- Confederate
  - Yes if a second person is using structures intended to prime the participant, but the participant is not aware
  - No otherwise

Note that many studies manipulate variables in addition to the moderators listed here, but including those is beyond the scope of this study.

Search strategies

The literature search was conducted using three primary methods: recording references listed in relevant review papers (V. S. Ferreira & Bock, 2006; Heydel & Murray, 2000; Pickering & Ferreira, 2008); searching for records which cite relevant work (Bock, 1986; V. S. Ferreira & Bock, 2006; Pickering & Ferreira, 2008); and searching ProQuest, Scopus, and Web of Science databases using natural language terms and controlled vocabulary. The third method was by far the most exhaustive, identifying 71 of the 73 papers included in the final list. The remaining two were found through the forward citation method (Goudbeek & Krahmer, 2012; Kantola & van Gompel, 2011). The first literature retrieval effort was conducted in July 2013. This yielded 70 of the total 73 records in the final list. A second retrieval was conducted in June 2015 in order to include records that had been updated since the first retrieval, up to the end of 2013. As a result, the final list of 73 includes all recovered records with journal publication dates prior to 2014 (2013 and earlier). Three additional papers were included in our initial analyses, but are not included in the final list of 73 papers: Biria, Ameri-Golestan, & Antón-Méndez (2010) which is about indirect questions and not strictly syntactic, Kootstra, Hell, & Dijkstra (2010) in which priming is only indirectly tested by comparing different experiments, and Shin & Christianson (2012) in which priming is used as a teaching aid making it not directly comparable to other production studies. An additional two papers in our final sample of 73 were added after a reviewer noticed the omissions; these two papers all appeared in the initial literature search and were erroneously excluded. For more details on the search procedure, see the supplementary material.

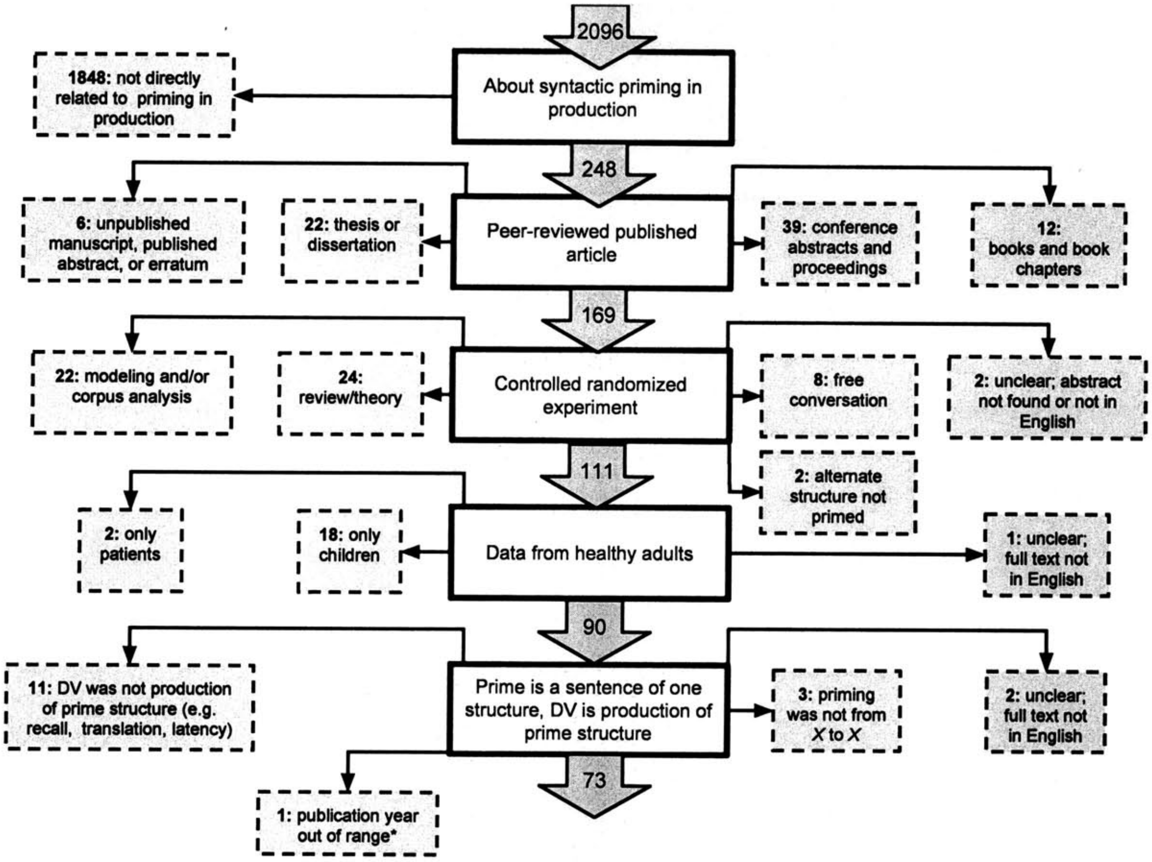

2096
About syntactic priming in production
1848: not directly related to priming in production
248
Peer-reviewed published article
6: unpublished manuscript, published abstract, or erratum
22: thesis or dissertation
39: conference abstracts and proceedings
12: books and book chapters
169
Controlled randomized experiment
22: modeling and/or corpus analysis
24: review/theory
8: free conversation
2: unclear; abstract not found or not in English
2: alternate structure not primed
111
Data from healthy adults
2: only patients
18: only children
1: unclear; full text not in English
90
Prime is a sentence of one structure, DV is production of prime structure
11: DV was not production of prime structure (e.g. recall, translation, latency)
3: priming was not from X to X
2: unclear; full text not in English
73
1: publication year out of range*

Figure 6-1: Flowchart showing literature search.

Coding procedures

From each experiment within each paper, we (K.M., R.F., and A.J.) extracted the number of unique subjects, the number of unique items, and the number of items per subject per experimental cell, along with the population characteristics and task characteristics needed for the moderator analysis. This coding was subsequently rechecked by K.M. For each experimental condition (e.g. verb repeated vs. not repeated), we extracted the mean proportion of productions matching the primed structure (e.g. prepositional-object dative; PO) and the alternative structure (e.g. double-object dative; DO). For a small number of papers, this information was not available. In these cases, the raw data (obtained from the original study authors) was used to obtain the estimates.

For every paper in the initial sample except the two added after the first round of peer review, one or more of the original authors was contacted by e-mail and asked for the raw data. Each author was contacted at least twice. For 45 of the 71 original papers, we received the raw data from the authors. For 25 of the 71 papers, the authors responded that the data was unavailable either because it was lost, corrupted, or otherwise inaccessible. For only 1 paper, we received no response from the authors.

Because information on "other" responses is not always available (both in raw data and in published estimates), our analysis excluded Other responses whenever possible. Thus, for all studies, the proportion of X responses (e.g., DO) and the proportion of Y responses (e.g., PO) add up to 1.

Statistical methods

To make meaningful comparisons across different studies, we need a uniform notion of "effect size" (Cohen, 1992; Lipsey & Wilson, 2001). For our purposes, we want the effect size to answer the question "how big is the effect of syntactic priming." To that end, the effect size measure we use is the log odds ratio of the prime condition compared to the no-prime condition (see Equation 1). That is, if the proportion of trials using the passive is .34 after a passive prime and .20 after an active prime, the log odds ratio would be `log(.34/(1- .34)) - log(.20 / (1 - .20))` = 0.72.[1]

(1) $\text{LogOddsRatio} = \log(\frac{p(X|\text{Prime})}{1-p(X|\text{Prime})})$ - $\log(\frac{p(X|\text{NoPrime})}{1-p(X|\text{NoPrime})})$.

In a meta-analysis, we do not want to give each study equal weight. Rather, studies which have smaller standard error (perhaps because of more subjects and items) should be weighted more. In order to know how much to weight each study in the meta-analysis, we need the standard error. We computed the standard error on the log odds ratio using the formula for standard error on a log odds ratio:

(2) $SE = \sqrt{\frac{1}{n_{PrimeX}} + \frac{1}{n_{NoPrimeX}} + \frac{1}{n_{PrimeY}} + \frac{1}{n_{NoPrimeY}}}$,

where n_{PrimeX} is the number of individual data points for which Structure X is primed and Structure X is used in the target, $n_{NoPrimeX}$ is the number of data points for which Structure Y is primed and Structure X is used in the target, n_{PrimeY} is the number of data points where Structure Y is primed and Structure Y is used in the target, and $n_{NoPrimeY}$ is the number of data points where Structure X is primed and Structure Y is used in the target.

When there was a baseline condition in addition to two prime conditions (i.e. DO prime, PO prime, and baseline), we ignored the baseline condition for our main meta-analysis (although it could be used in the p-curve analysis). We excluded studies for which either of the condition means was above .98 or below .02 or in which both condition means were either above .90 or below .10, since the log odds ratio is inflated near 0 and 1. In addition to being problematic for the quantitative analysis, we do not believe that these studies are directly comparable to the studies in which participants are less categorical in their use of particular constructions. Results from 43 experimental conditions of an original 386 data points were excluded for this reason, including all the data from 3 papers. We also excluded Coyle & Kaschak (2012) since item and prime condition are confounded in that experiment, and thus it is not possible to obtain an estimate of the size of the priming effect.

[1] We use the log odds ratio instead of the odds ratio because it has better distributional properties, but it can easily be converted to an odds ratio by exponentiating.

These methods for computing log odds and standard error are not necessarily optimal given the within-subject, within-item designs common in psycholinguistics. The current standard for estimating the effect size and standard error in categorical data like this, given the latest statistical thought (Barr et al., 2013; Bates et al., 2015; T. Jaeger, 2008), is to extract size and its associated standard error from a linear mixed effect logistic regression with random effects for subject and item. But this information is not always available in published reports for a variety of reasons. First, before 2008, researchers typically reported ANOVAs instead of mixed effect regressions and the reported results of those ANOVAs are not usually sufficient for computing standard error that is comparable to a standard error from a mixed effect regression. Moreover, in both mixed effect regressions and ANOVAs, there is variability as to how the random effects are structured across papers that may make them incomparable. And often the hypothesis being tested is not just whether priming exists but about some other variable–in which case sufficient test statistics for the actual priming manipulation may or may not be reported.

Thus, we believe that the best ways to ensure that estimates are consistent across papers is to a) use raw trial-level data (which we obtained for a subset of papers) and analyze all experiments together in one model and b) use the published means and design characteristics (which are almost always available) to compute the log odds ratio and the standard error. We used both of these techniques and, as we report below, found similar results using both methods.

6.2.2 Results

Characteristics of studies remaining

From our initial 73 papers, we analyzed a total of 343 data points (i.e., experimental conditions) from 138 experiments from 69 papers. The median number of participants per condition was 32. The median number of items seen by each participant in each experimental cell was 6. The full list of included studies is given in the appendix. Summary statistics (unweighted mean effect size and number of data points per moderator) can be found in Figures 6-2 and 6-3.

Weighted mean results

To facilitate meaningful comparisons across studies, we computed a weighted mean effect size as described in Lipsey & Wilson (2001). The clearest and most consistent moderator of the size of the priming effect was lexical overlap between the prime and target (i.e., whether the same word, a semantically or a phonologically related word, or a translation-equivalent word was repeated in the prime and the target). For 220 studies with no lexical overlap, the weighted mean odds ratio was 1.67 with a 95% CI of [1.63, 1.72], $p < .0001$, such that the odds of a construction occurring are 1.67 times greater when it is primed than when it is not primed. This means that, if a construction occurs 50% of the time when it is not primed, it would occur 63% of the time when primed. Multiplying the log odds ratio by $\frac{\sqrt{3}}{\pi}$ to convert it to an estimate of a Cohen's d standardized effect size (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hasselblad & Hedges, 1995), estimated d is 0.28 (a small-to-medium-sized effect per Cohen's original rubric (J. Cohen, (1977)).

There were 123 studies with lexical overlap. The weighted mean odds ratio was 3.26 with a 95% CI of [3.13, 3.40], $p < .0001$, such that the odds of a construction occurring are 3.26 times greater when it is primed than when it is not primed. If a construction occurred 50% of the time when not primed, it would occur 77% of the time when primed with lexical overlap. Converting the odds ratio to Cohen's d, we estimate an effect size of d = 0.65 (a medium-to-large-sized effect).

We show means and 95% CIs for the studies and conditions in Figures 6-4-6-6. These means and standard errors are based on the published estimates for each paper.

A simple weighted mean does not account for additional structure in the data, such as the correlations between conditions of the same experiment, the correlation between experiments in the same paper, and the various moderators of syntactic priming that are manipulated within and across experiments. Therefore, we next present results from a mixed effect meta-regression.

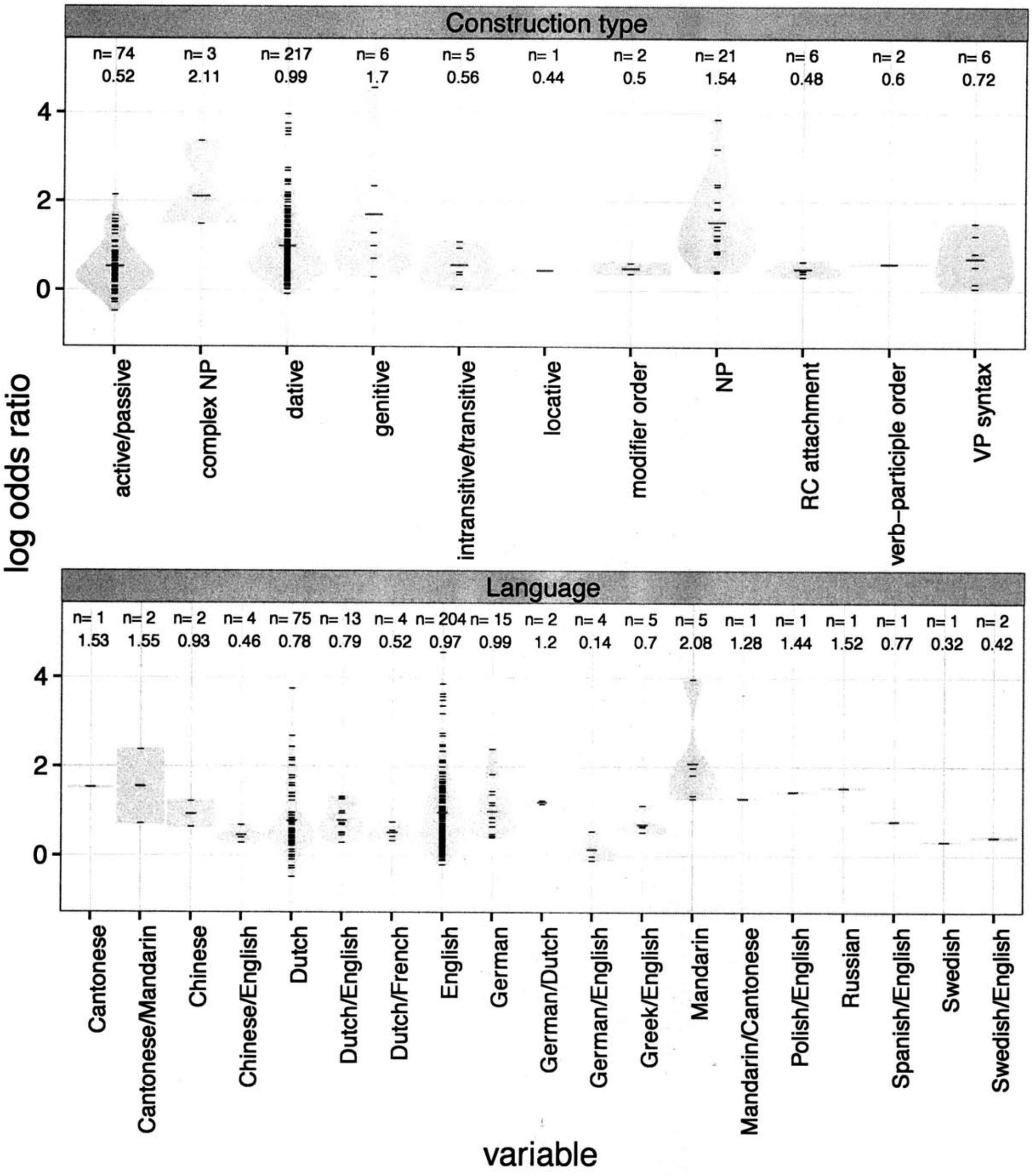


Figure 6-2: Effect size estimates (one data point per experimental condition) by language and construction type are represented by the individual horizontal lines and are not weighted by sample size or standard error. The horizontal line represents the median, and the gray blobs represent smoothed density estimates such that fatter parts of the blob represent more likely value.

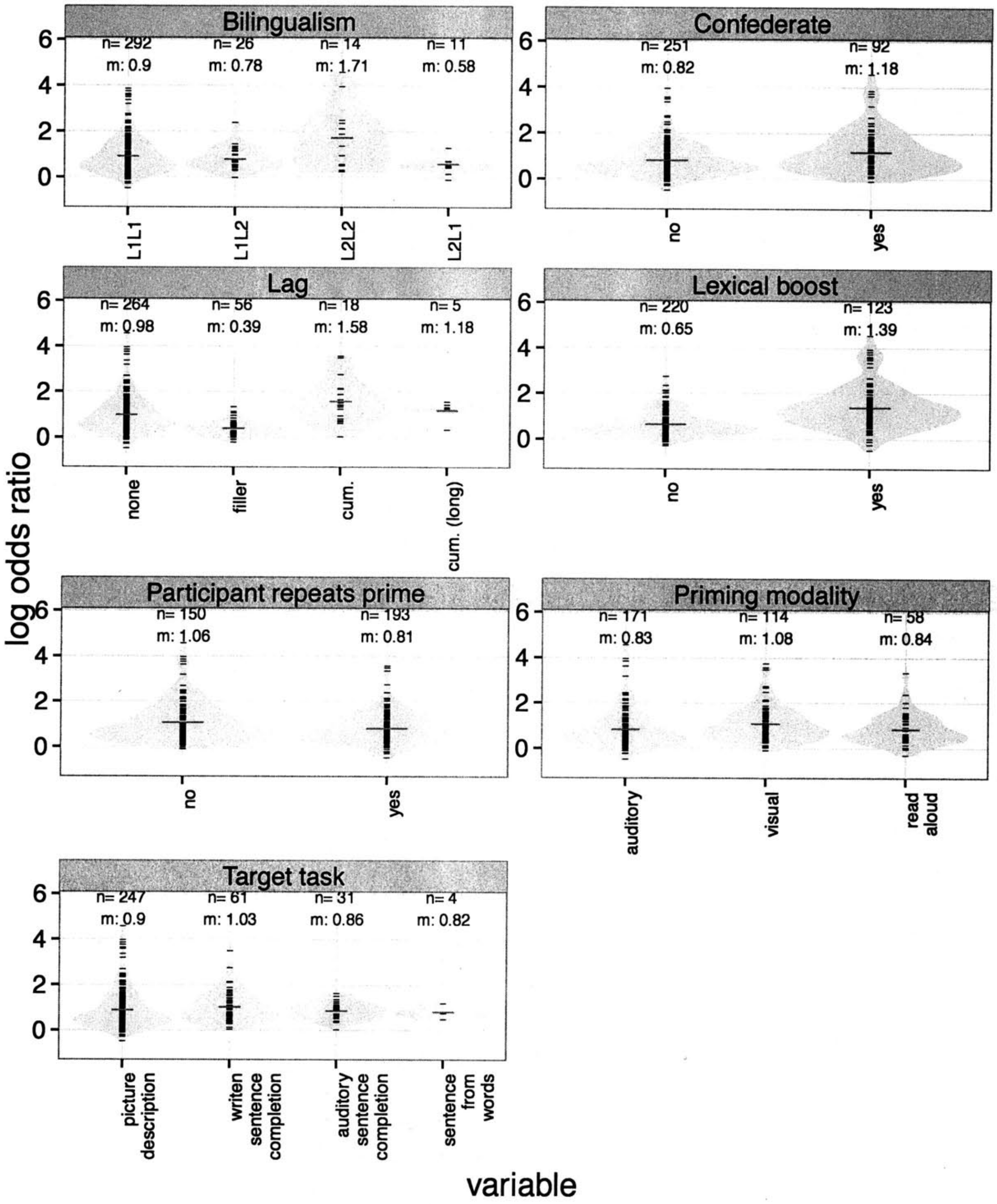


Figure 6-3: Effect size estimates (one data point per experimental condition) by moderator are represented by the individual horizontal lines and are not weighted by sample size or standard error. The horizontal line represents the median, and the gray blobs represent smoothed density estimates such that fatter parts of the blob represent more likely value.

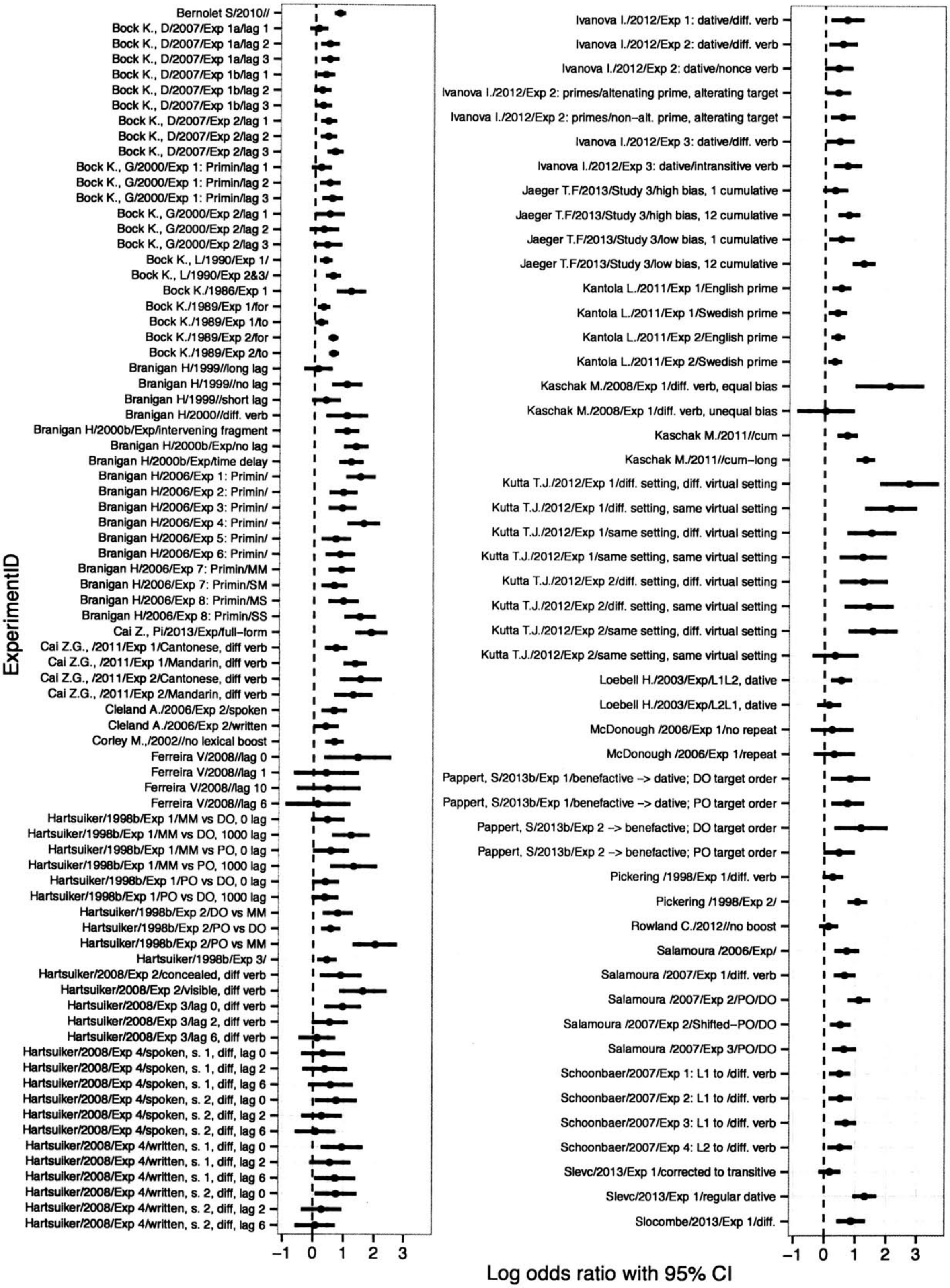


Figure 6-4: Forest plot with 95% CIs for dative studies with no lexical overlap

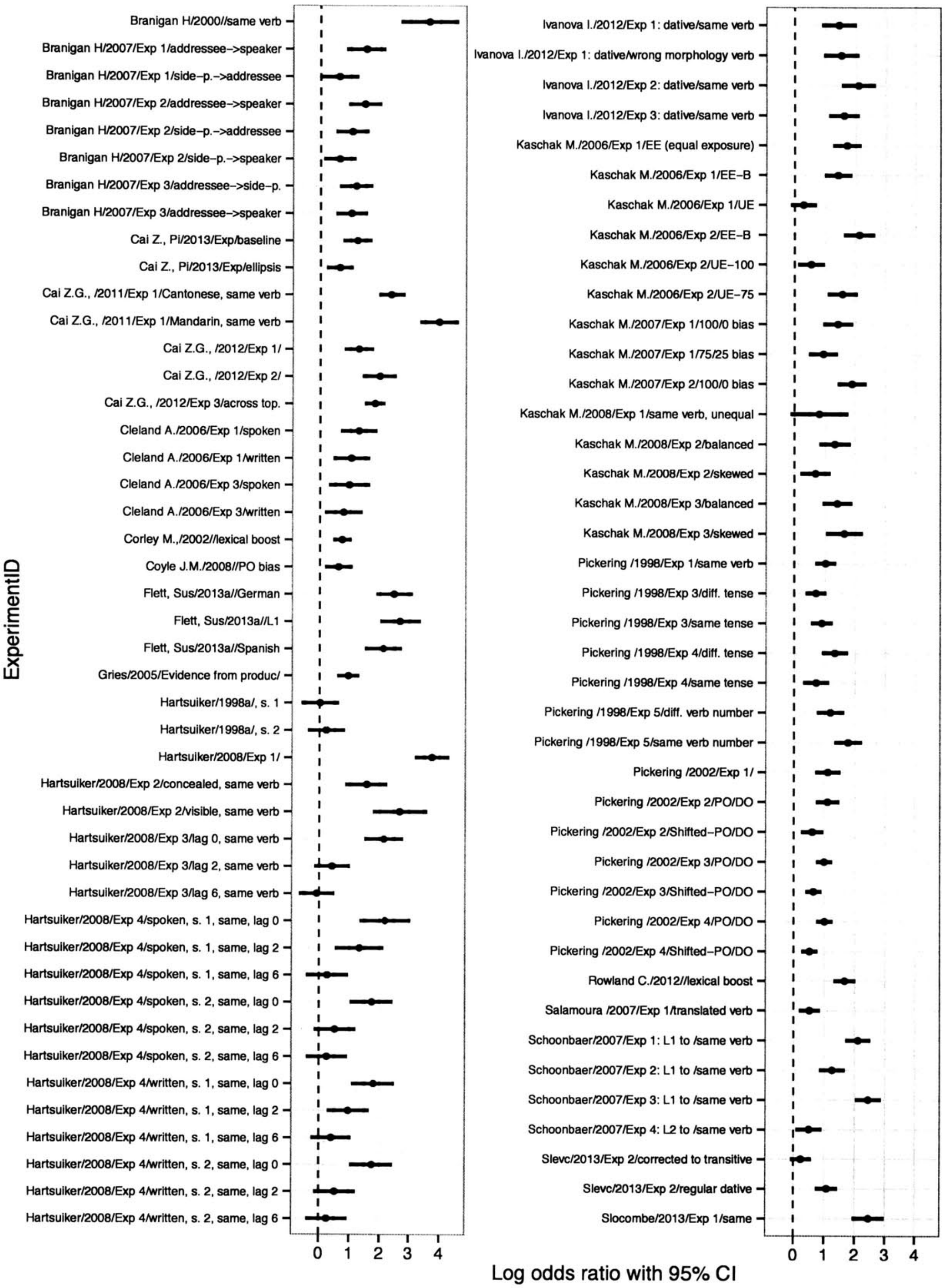


Figure 6-5: Forest plot with 95% CIs for dative studies with lexical overlap

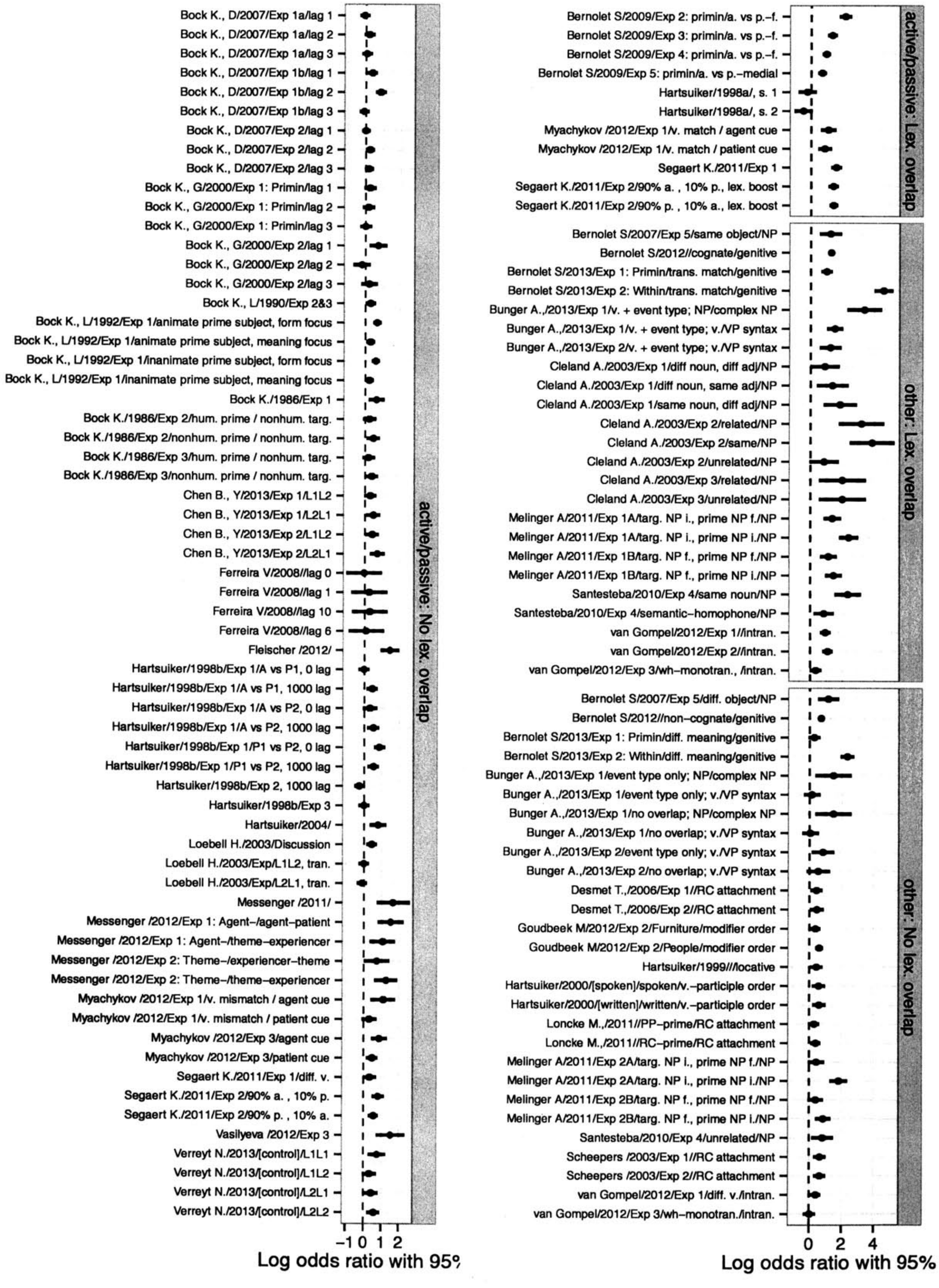


Figure 6-6: Forest plot for active/passive and other studies with and without boost.

Model results

We first fit a random effect, intercept-only meta-analysis model to all data points and found a significant intercept of 0.87 [95% CI 0.80, 0.94], $p < .0001$. This indicates a significant effect of syntactic priming in our sample. Since these studies sometimes include very different experimental conditions, there was unsurprisingly significant heterogeneity in this estimate, as measured by a Q-test comparing the variability among effect size estimates to the expected sampling variability: Q(342)=2748.60, $p < .0001$.

None of the estimates reported above account for moderators of syntactic priming. For instance, perhaps priming exists for certain constructions but not others. Using the `metafor` package (Viechtbauer, (2010) in `R` (R Core Team, (n.d.), we fit a mixed effect meta-analysis regression to the data. The mixed effect meta-analysis differs from a standard mixed effect model in that the standard error of each data point is assumed to be known instead of estimating it from the data. But the underlying logic is the same in that we are asking what underlying parameter values could plausibly give rise to the observed effect sizes obtained in the published studies. Here, each individual data point is an effect size (change in log odds ratio) extracted from an experimental condition with an associated standard error (where standard error is estimated as described above, using the number of subjects, items, and the condition means).

In this meta-analysis regression, the intercept is the size of the priming effect. We included fixed effects of lag, year, lexical overlap between prime and target, within/between language condition, target task, mode of presentation, whether the participant repeated the prime, and whether there was a confederate. Since it has been widely posited in the literature that lexical overlap interacts with temporal lag and that lexical overlap exerts a stronger effect on priming in L2 populations, we included interactions between lexical overlap and lag and between lexical overlap and bilingualism. To avoid overparameterizing the model, we did not include further interactions–especially since few papers vary by more than one factor and thus the interaction terms would be ill-defined. We included random effects for construction type (dative, active/passive, etc.), paper, experiment (nested within paper), and condition (nested within paper and experiment). We did not include random slopes since none of these grouping factors consistently vary by anything but prime condition.

Across 69 papers that survived all exclusions, consisting of 138 experiments and 343 unique conditions, we found a significant baseline priming effect (no lexical overlap, no lag) corresponding to a change in odds ratio of 1.68 [95% CI 1.25, 2.27]; Cohen's d = 0.29 when there is no lexical overlap between prime and target and 3.67 [95% CI 2.53, 5.31]; Cohen's d = 0.72 when there is lexical overlap between prime and target. Thus, as with the simple weighted mean analysis, the model suggests that the size of the priming effect is small-to-medium without lexical overlap between prime and target and medium-to-large with lexical overlap.

Moderators

The moderators are shown with their estimates and 95% CI's in log-odds space in Figure 6-7. Terms significant in either analysis at $p < .05$ are starred, with two stars for $p < .01$ and three stars for $p < .001$. The intercept in this model is a priming study with no lexical boost, no lag, no confederate, an auditory picture description task with an auditory prime that is not repeated by the participant, performed in the year 2000. Besides the main effect of priming, the coefficients in Figure 6-7 represent the change in log odds ratio associated with adding that moderator. As a rough rule of thumb, Gelman & Hill (2007) suggest dividing by 4 to convert the log odds coefficients to changes in probability space.

The presence of the moderators significantly reduces the heterogeneity in the data: QM(20) = 230.01, $p < .0001$. Even with the moderators included in the model, though, there is still significantly more variance than expected by sampling variability alone: Q(305) = 1490.53.

Lexical overlap between prime and target Lexical overlap is the most consistent moderator of syntactic priming (this is the "lexical boost" effect first demonstrated in

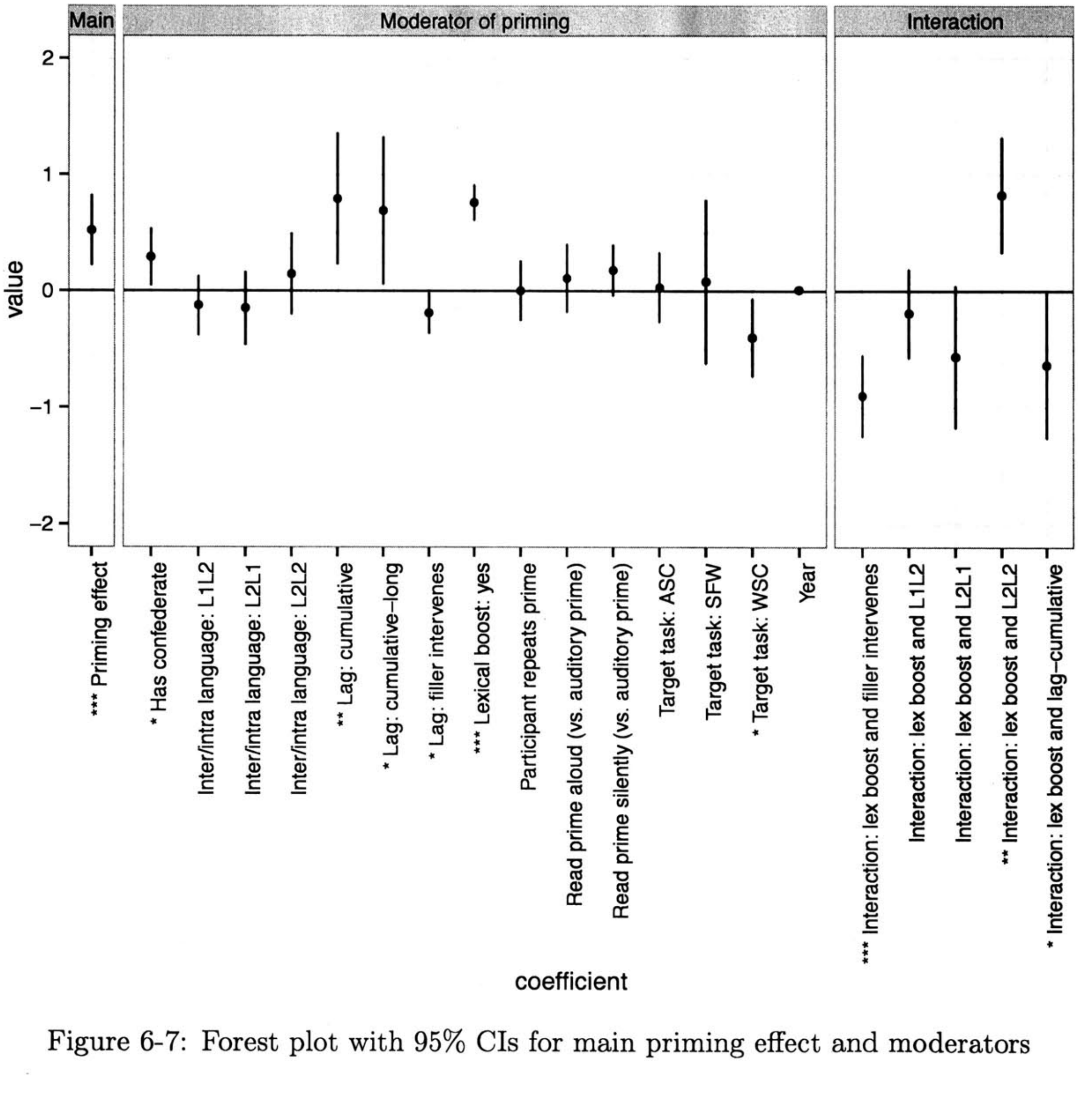


Figure 6-7: Forest plot with 95% CIs for main priming effect and moderators

Pickering & Branigan (1998)). It significantly enhances the priming effect (beta=0.76, z=9.9, p<.0001), and the effect of lexical overlap is actually stronger than the priming effect itself (i.e., the change in participant response tendency between priming without lexical overlap and priming with lexical overlap is greater than the change from no priming at all to priming). See Pickering & Ferreira (2008) for a clear summary of the many reasons that have been proposed for the strength of the lexical overlap effect.

Modality of prime Mode of prime was analyzed in three separate coefficients: modality of prime, whether the participant generates the prime (either by having to produce it herself or by simply rotely repeating it), and whether there was a confederate. The baseline here was an auditorily presented prime that was not repeated, with no confederate. We found no clear effect of modality. Although we did not test cross-modal priming since only a small number of papers in our sample studied it, we take this to be broadly consistent with findings in the literature (Cleland & Pickering, 2006; Hartsuiker & Westenberg, 2000) that the modality of the prime does not strongly affect the size of the syntactic priming effect. There was also no clear effect of having a confederate.

Target task In our model, the baseline target task is the classic picture description task (Bock, (1986). Consistent with Hartsuiker & Westenberg (2000), relative to picture description, auditory sentence completion (Branigan, Pickering, Stewart, & McLean, 2000; Hartsuiker & Westenberg, 2000) produced similar-sized priming effects, as did the task where participants are asked to generate sentences from a list of words (although only one paper, Pappert & Pechmann (2013), used that strategy so little can be concluded about it besides what is concluded there). We found a marginally significant negative effect for written sentence completion (Pickering & Branigan, (1998), such that written sentence completion results in less priming than picture description. Every written sentence completion task in our sample, however, does find a numerically positive priming effect.

Bilingualism We consider three types of bilingual priming: priming within a second language, priming from a first language to a second language, and priming from a second language to a first language. We did not find significant main effects for priming to be stronger or weaker for bilingual priming relative to classic priming within a native language, although cross-linguistic priming was numerically weaker than within-language priming. This is consistent with Pickering & Ferreira (2008), who suggest that the effect of cross-linguistic priming is similar to that of L1-L1 priming. The trend towards priming being weaker cross-linguistically may be, in part, driven by the inclusion of cross-linguistic priming effects in which the languages used have different word orders. For instance, Bernolet et al. (2007) found little to no priming between Dutch and English for complex noun phrases, possibly since the structures are so different as to be represented differently.

Importantly, we found that, relative to L1→L1 priming, there was a strong enhancement of the lexical overlap effect when priming took place in a second language (beta=0.83, z=3.27, p<.01). This is consistent with past results (e.g., Kim & McDonough (2007), Schoonbaert et al. (2007), Bernolet et al. (2013)), which suggest that L2 speakers are highly sensitive to lexical boost and maintain strong item-specific representations. Specifically, Bernolet et al. (2013) found that less proficient speakers are more susceptible to lexical boost effects–perhaps suggesting that less proficient speakers rely on item-specific representations.

We also found a significant trend for the lexical overlap to be smaller for L2→L1 priming than for the L1→L1 baseline (beta=-0.57, z=-1.82, p= 0.07). This is perhaps not surprising since, in the former case, "lexical overlap" typically refers to a translation-equivalent word, whereas in within-language priming it is often an identical word. There was a smaller, not significant trend for L1→L2 priming (beta=-0.19, z=-1, p= 0.32) to also show less of a lexical boost effect than within-language priming. These findings–an asymmetry in lexical boost between within-language and between-language priming and a further assymetry in lexical boost bewteen L1→L2 priming and L2→L1 priming–are consistent with what Schoonbaert et al. (2007) found and used to argue in favor of the lexical-syntactic model described in Hartsuiker, Pickering, & Veltkamp

(2004).

Lag We considered three types of lag: the inclusion of fillers between the prime and target, cumulative priming, and long-term cumulative priming (more than one day between prime and target). The latter two are techniques developed and used mostly by Kaschak and colleagues (e.g., Kaschak (2007), Kaschak & Borreggine (2008)) whereby a group of prime sentences are presented in a block (priming either a construction or a particular construction with a particular verb), which is then followed by a target phase. While we treat this as a type of lag, temporal delay is not the only difference between these paradigms and other syntactic priming paradigms.

The baseline condition here is no temporal lag between prime and target. There has been some debate in the literature as to how much, if at all, temporal lag reduces the priming effect. We found a small but significant negative main effect of including filler material between prime and target (beta=-0.19, z=-2.17, p<.05), such that the priming effect was smaller when there was filler material. We also found that including one or more fillers between the prime and target significantly reduces lexical boost (beta=-0.9, z=-5.08, p<.0001). This reduction of lexical boost is strikingly large relative to the main effect of lexical boost and the main effect of filler, suggesting that the lexical boost effect essentially disappears entirely when there is filler material between prime and target. Hartsuiker et al. (2008) found something similar and argued that it reflects evidence for both a long-term implicit learning account of syntactic priming as well as a short-term lexical component.

There was a strong main effect for priming to increase relative to the no-lag standard priming condition when the paradigm uses cumulative priming (beta=0.79, z=2.75, p<.01) or long-term cumulative priming (beta=0.69, z=2.13, p<.05). But, as with filler lag, the effect of lexical overlap is reduced using this paradigm (beta=-0.64, z=-2.01, p<.05). Note that cumulative lag, in our meta-analysis, is largely confounded with primes that are presented in blocks as opposed to as single sentences. So the increased effect of priming using this paradigm, as seen in the large main effect, is plausibly the result of the paradigm as a whole and not *because* of the temporal delay between the primes and the target.

Year of publication Centering year at 2000, we found no effect of year of publication on the size of the priming effect (beta=0.01, z=1.2, p= 0.23).

Validating the model using raw data

To test whether the meta-analysis regression described above is appropriate and whether the effect sizes and standard errors used are good estimates of the data, we fit a mixed effect logistic regression to the 71194 trial-level data points from the 43 non-excluded papers for which we have raw data (2 were excluded for reasons described above).

For each experiment, we set the dependent variable to be the less frequent of the two syntactic constructions and predicted the dependent variable from fixed effects of prime condition and the interactions of prime condition with the following:

We used the same fixed effects as for the meta-regression above, with the exception that the intercept here is the response variable and thus there is a main effect of prime condition. Moderators are represented not as main effects but as interactions with prime condition. But the underlying logic is the same, and we will treat these interactions straightforwardly as moderators of the priming effect. We included random effects for construction type (dative, active/passive, etc.), paper, experiment (nested within paper), condition (nested within paper and experiment), unique subject, and unique item. Subject and item are necessarily nested within experiment, and although it is likely that some subjects participated in multiple experiments and even more likely that some items were re-used across experiments, we do not have data to model this. Following Bates et al. (2015) and using the `RePsychLing` package, we first fit the maximal justified model and then simplified it to obtain convergence and avoid overparameterization. We found that including the correlation parameters did not significantly improve the model. We included random slopes for prime condition by subject, item, experiment, and condition. Other random slopes did not significantly improve model fit or led to lack of convergence.

We compared the results of this regression to a meta-regression on the same set of data (i.e. just the subset of experiments for which we have raw data). We found that the results are qualitatively and quantitatively similar to those obtained using the mixed effect model on the raw data and that, for most parameters, the 95% confidence interval includes the point estimate from the other method. Figure 6-8 shows the fixed effect coefficients with 95% CIs from the model for this data from the raw data regression (in gray) and the meta-analysis regression (in black). We thus believe that the meta-analysis technique used here, based off published estimates, is sufficient to give similar results as having all the published data.

Having said that, the meta-analysis regression using all the data gives an odds ratio (1.68) that is higher than the margin of error of the same regression run on just the subset of studies for which we have raw data (1.46). While it is possible there is systematic difference between the studies for which we were able to obtain raw data and studies for which we were not, the estimates are similar enough that the difference is likely just noise.

Interim conclusions

Using the published estimates of the priming effect across 69 papers, we found a robust effect of syntactic priming that becomes dramatically larger when there is lexical boost. We also found effects of temporal lag, bilingualism, and year of publication. We validated the model on a subset of papers for which we have raw data and found similar results.

In any meta-analysis, though, the conclusions that can be drawn are only as sound as the data that goes into it. In our case, our meta-analysis was restricted to only published studies in peer-reviewed journals.[2] As a result, to assess the validity of our results, we must ask whether the sample suffers from low statistical power, publication bias, or p-hacking. If the only articles that are accepted to journals are ones that include significant results or if study authors performed multiple analyses and only

[2] We also solicited unpublished studies, but we received only a handful of responses including studies that were not likely to be published in the next year. Therefore, we did not include these in the current report.

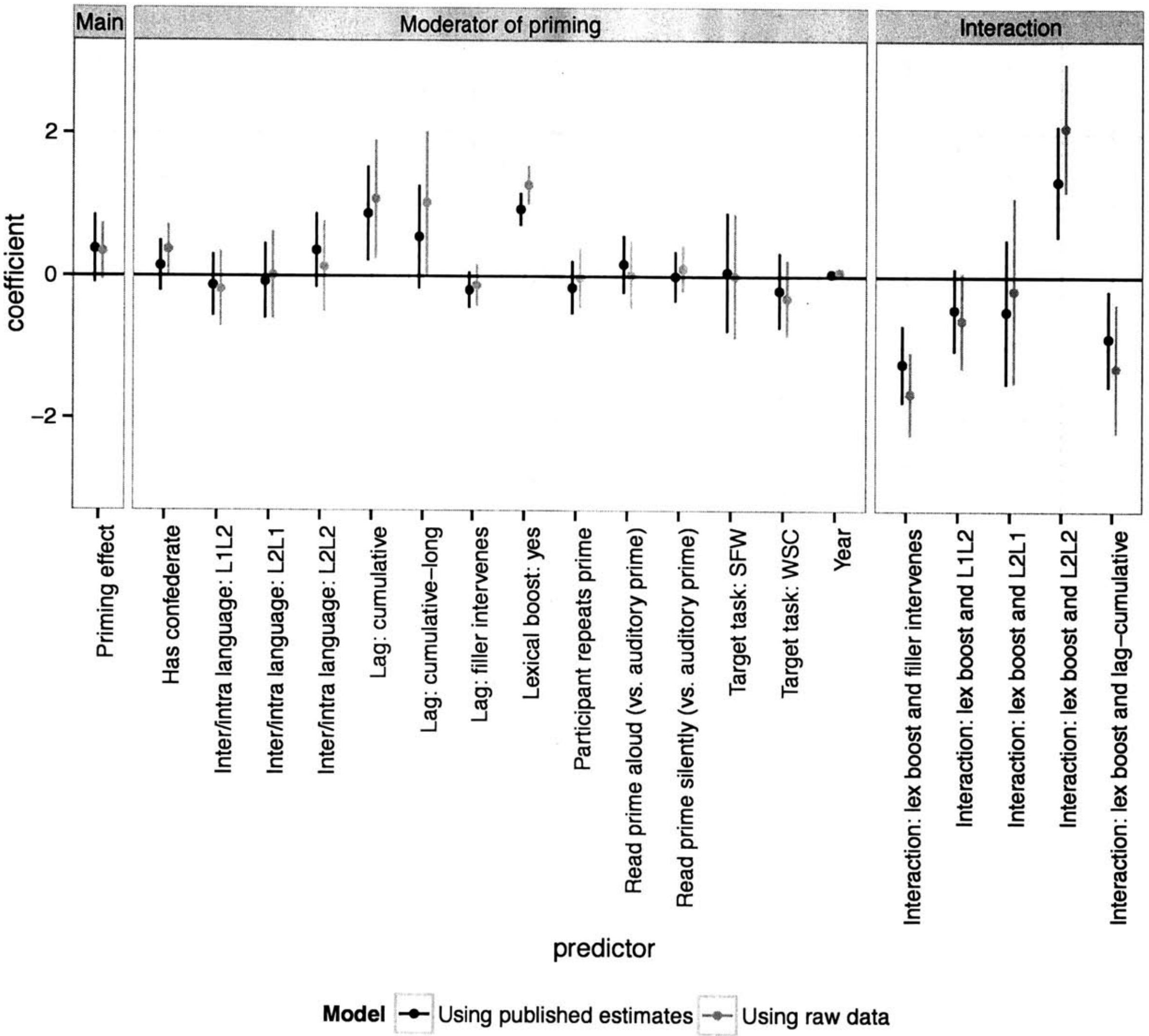


Figure 6-8: Forest plot with 95% CIs for main priming effect and moderators

reported the significant ones, the effect sizes here would be inflated. In the next section, we analyze the distribution of p-curves in the studies sampled in order to evaluate the evidential value of the results

6.3 Assessment of publication bias and statistical power

The distribution of p-values used to support or refute the hypotheses of a particular set of studies can be used to test for evidence of publication bias or p-hacking in those studies (Francis, Tanzman, & Matthews, 2014; Simonsohn et al., 2014a, 2014b). Recall that a p-value is the probability of a null hypothesis having generated data as extreme as the data observed. In psychology studies, p-values less than .05 are taken as sufficient evidence to reject the null hypothesis. In studies investigating effects of syntactic priming, the null hypothesis is usually that there is no difference between the two prime conditions. If a study is investigating whether or not there is lexical boost, the null hypothesis would be that lexical overlap between prime and target has no effect on the strength of priming.

If a group of studies is investigating a real, robust effect, the distribution of p-values will be right-skewed such that there are more p-values between 0 and .01 than between .02 and .03, more p-values between .02 and .03 than between .03 and .04, and so on. Just how sharply skewed the "p-curve" is will be a function of statistical power: high power leads to more right skew. This is perhaps simplest to think about in the extreme case. With infinite sample size and a true effect (even if that effect is small), the p-values would be arbitrarily small (certainly all less than .01). If a group of studies is investigating a *null* effect, though, the p-values will be uniformly distributed between 0 and 1. Now imagine that only p-values less than .05 were published in journals and so we only have access to p-values less than .05. If the underlying effect is real, we will still see a right-skewed p-curve. If the underlying effect is *not* real (i.e., the distribution of p-values for experiments is uniform between 0 and 1), when we censor all p-values greater than .05, we will be left with a flat distribution of p-values between 0 and 1. Note that, if statistical power is low (i.e., the probability of correctly rejecting the null when the null is false is too small), that could also lead to a flat distribution of p-values. In either case, a flat p-curve would suggest a lack of evidential value in the data. If researchers are actively p-hacking–that is, re-running variants of analyses until a significant result (p<.05) is obtained, then the curve will actually be left-skewed such that there are more p-values between .04 and .05 then between .03 and .04.

The goal of our p-curve analysis is to assess whether the collection of experiments identified in the meta-analysis contain evidential value for the claims they make. The papers in our sample make two fundamentally different types of claims, however. Some argue that syntactic priming exists or does not exist given various experimental conditions. Others argue that some moderator significantly affects the size of the syntactic priming effect. To that end, we made a pre-analysis decision to split the studies for p-curve analysis into two groups to be analyzed separately: those in which the main effect of interest was a main effect of syntactic priming and those in which the main effect of interest was a moderator of syntactic priming (e.g. whether using a temporal lag impacts the syntactic priming effect).

6.3.1 Method

Inclusion and exclusion criteria for p-curve

In p-curve, we can use only one p-value from each experiment. To decide which p-value to use, we identified the main statistical prediction of each study. This step is important since the p-curve analysis critically depends on the fact that the set of p-values included in the p-curve actually test the experimental hypothesis. Including auxiliary p-values that do not actually relate to the hypothesis of the study could cause the p-curve to be uninformative (Simonsohn et al., 2014a). Consequently, studies whose central claim was not about syntactic priming were not included in the p-curve analysis. Studies which predicted and/or found null results can also not be included in p-curve. Our sample included many such experiments.

As an example, consider Cleland (2003) Experiment 2. This study investigated priming of noun phrases in three conditions: when the noun is the same in the prime and target (sheep/sheep), when the noun is unrelated in the prime and target (knife/sheep), when the noun is semantically related in the prime and target (goat/sheep). The question was whether the priming effect is moderated by the relationship between the prime noun and target noun. One reported result is that there is a lexical boost effect such that there is an interaction between prime and the 3-condition factor semantic relatedness. But this result could be driven by simple lexical boost since it includes the same noun condition vs. the different noun condition. The main prediction of theoretical interest–and the apparent raison d'etre for the experiment–is that the semantically related condition will differ from the unrelated condition. Thus, the p-value that we p-curve is the p-value for the planned comparison between the semantically related condition and the unrelated condition. The fact that this is the main claim of interest is made clear in the abstract, in which the contribution of Experiment 2 is distilled to one sentence: "Experiment 2 showed an enhanced priming effect when prime and target contained semantically related nouns (e.g., 'goat' and 'sheep')."

Among the studies that were included in our main meta-analysis, we used the following criteria to determine inclusion or exclusion of p-values in the p-curve analysis.

- We excluded result in which the statistical result appears across two or more experiments in the same paper. That is, one data point in p-curve corresponds to one experiment and cannot be a combined analysis of Experiments 1 & 3 in a paper.

- We excluded experiments in which the main claim involves a comparison with a population excluded from our meta-analysis (i.e., populations like aphasic patients or children). (For these studies, the control groups may appear in our main meta-analysis, but typically there is no main claim being made about how the control group will behave.)

- We excluded experiments in which the only main claim involves a dependent variable other than elicited sentence production (e.g., eye-tracking or reaction time). If an experiment has multiple main claims, some of which are about elicited sentence productions, we will include only the claim about elicited sentence production.

- When we cannot determine a clear "main result", we recorded the p-values for the main results in the order in which they appear (henceforth known as results "a", "b", "c", etc.).
- When an experiment reports an ANOVA by subject and item (F1 and F2) or any other analysis that analyzes subjects and items separately, we take the higher p-value of the two produced since using just one or the other is massively anti-conservative (Barr et al., (2013) and thus violates an assumption of p-curve: that the p-value reflect the actual false-positive rate of the statistical test. Moreover, since the criteria for significance is that both F1 and F2 give p-values less than .05, any p-hacking would take place on the higher of the two p-values.

Coding procedures for p-curve

For the p-curve analysis, we created a p-curve disclosure table (available at `https://osf.io/b9zyk/`), following best practices (Simonsohn et al., 2014a, 2014b). One of the authors (K.M., A.J., R.F., or E.G.) first coded the results, and they were all independently recoded by another author. K.M. then went over each result in which the coder and re-coder's results did not match and discussed them. In cases of obvious error, the correct version was kept. In cases in which there was legitimate disagreement as to which p-value best reflected the main hypothesis of an experiment, both versions were used and the re-coded version was used as a "b" result.

Statistical procedures for p-curve

Following the procedures in Simonsohn et al. (2014), observed p-curves are tested for significant right skew (most p-values near zero) using Stouffer's method. We also compared the observed p-curve to a hypothetical p-curve with 33% power. Simonsohn et al. suggest that observed curves that are flatter than the 33% power curve suggest a lack of evidential value. Further comparisons with curves of varying statistical power are the basis for the estimated average power of the included studies. Essentially, we ask what power level is most likely to produce the observed curve shape. See Simonsohn et al. (2014) for more details of the p-curve analysis.[3]

To see how robust the p-curve analysis is to different plausible decisions about how to code the papers, whenever there were multiple results coded, we randomly sampled one per experiment. We did this 100 times for each experiment to represent 100 possible coding procedures and used the average as an estimate of power.

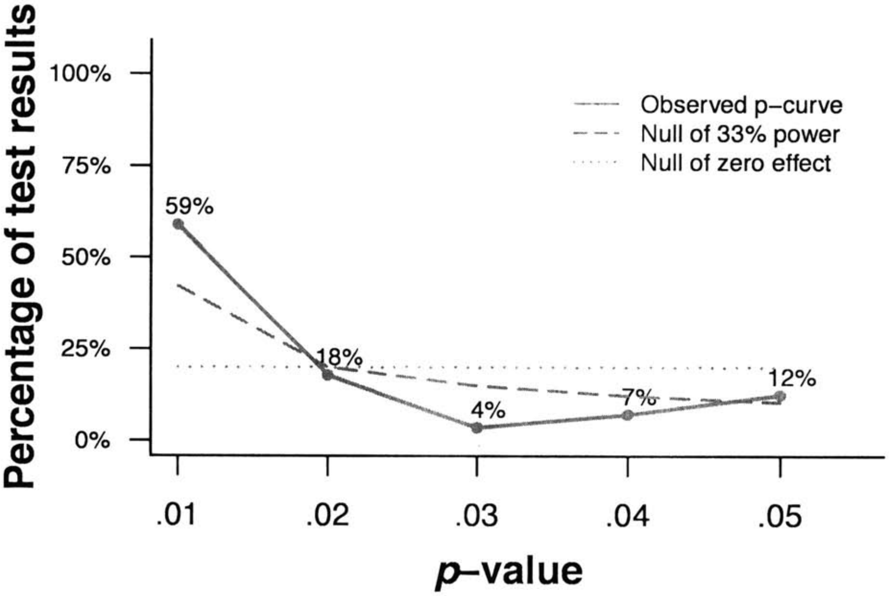


Figure 6-9: Default plot produced by p-curve for sample of studies in which the effect of interest is the existence of syntactic priming. The blue line is the distribution of p-values in the study. The red line shows the expected distribution of p-values if there was no underlying effect. The green line shows the expected p-curve under 33% statistical power. The right skew in the blue line shows there is evidential value in this set of studies.

[3] Note that Simonsohn et al. (2014) cast doubt on the validity of p-curve for discrete tests since the p-value from a discrete test does not necessarily correspond to the false-positive rate. In our case, while we are ultimately concerned with a comparison of proportions, the p-values used in psycholinguistics are typically designed such that the p-value reflects the actual false-positive rate (Barr et al., (2013). Therefore, this should not be a major issue.

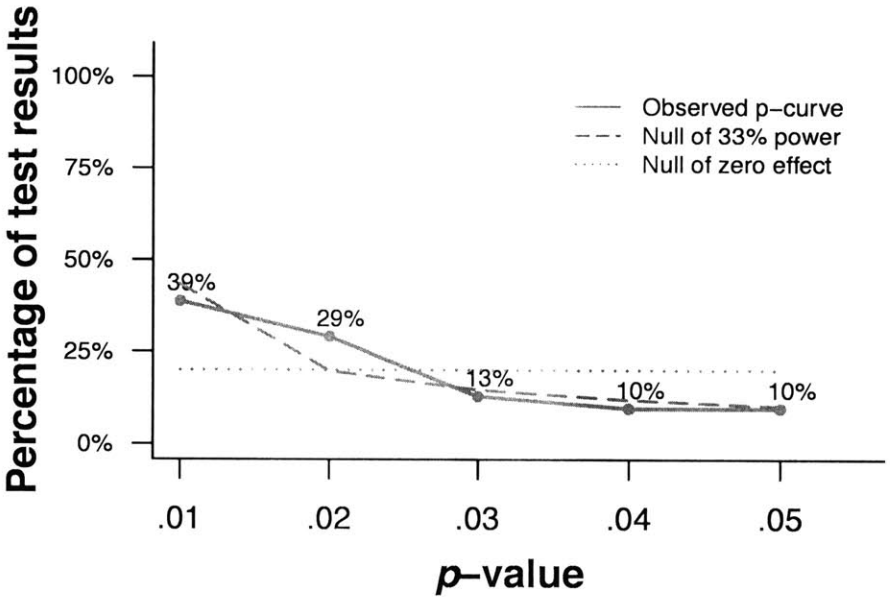


Figure 6-10: Default plot produced by p-curve for sample of studies in which the effect of interest is a moderator of syntactic priming. The blue line is the distribution of p-values in the study. The red line shows the expected distribution of p-values if there was no underlying effect. The green line shows the expected p-curve under 33% statistical power. The right skew in the blue line shows there is evidential value in this set of studies.

6.3.2 P-curve results

After eliminating studies in which the main hypothesis was about excluded populations or in which there was no priming-related main hypothesis internal to a single study, we were left with 139 studies from 65 papers. Of those 139 studies, only 88 (63%) had a significant result used to support the main hypothesis.

Figure 9 plots the distribution of p-values for a representative choice of what the "main" hypothesis is from the set of papers, for 56 experiments that directly test the existence of some form of syntactic priming, in blue. As recommended by the p-curve guide, the green line shows the expected p-curve under 33% power. We see that most studies in our sample (59%) have p-values less than .01, and the curve shows significant right skew ($p < .0001$ by a Stouffer's z-test for skew). Comparing the blue line to the green line shows that there are many more small p-values than we would expect if the power were 33%. The bias-corrected average power estimate is 77%, which is near the recommended 80% threshold (a minimum standard for statistical power in many fields). There are numerically more p-values at the .04 and .05 levels than at the .03 levels (which is not consistent with a healthy p-curve), but the distribution is not extreme enough to conclude that there is definitive bias.

In a post-hoc p-curve analysis that we ran after a reviewer noticed that the recommended sample size for 80% power in priming studies with no lexical overlap was higher than the sample size of most papers in our sample, we ran a separate p-curve power analysis on just a subset of experiments (n=32) that contain no lexical overlap between prime and target. (If an experiment contained overlap and non-overlap conditions, it was not included in this subset.) For those papers, we found that average bias-corrected power was only 54% [32%, 72%]. This analysis suggests that, in the absence of lexical repetition between prime and target, studies in our sample may be underpowered.

Figure 10 plots the distribution of p-values for 32 experiments that do not directly investigate the existence of syntactic priming but ask questions about how syntactic priming is moderated by other variables. The p-curve for these studies is quite a bit flatter, although still significantly right skewed ($p < .01$). Although publication bias cannot be ruled out as an explanation for the flat p-curve, the p-curve is consistent with low statistical power (see 33% power curve, plotted in green, for comparison; p-curves significantly flatter than this line can be taken as evidence that the included studies lack evidential value). While the estimated power for these studies is only 53% [32%, 71%], the right skew and the fact that the curve is not significantly flatter than the 33% power curve ($p = .96$) suggests these studies do contain evidential value.

Interim conclusions

These results suggest that studies which purport to directly investigate the existence of syntactic priming do not suffer from extreme p-hacking and are moderately powered. Studies which investigate moderators of syntactic priming are likely underpowered but still contain some evidential value. The former suggests that our main meta-analysis is likely not overly influenced by publication bias or p-hacking.

One possible reason that studies investigating moderators of priming are underpowered is that these designs are more likely to involve interactions. Studies like these typically need many more subjects or items than those that are simply investigating a main effect of priming. In the next section, we will investigate just how many subjects and items need to be included to run well-powered syntactic priming studies of varying design.

Overall, the p-curve results should be interpreted with caution. Each p-curve analysis aggregates over a heterogeneous group of studies. In the p-curve analysis of studies that directly test priming, for instance, there are both studies with lexical overlap between prime and target and studies without. The ones with lexical overlap have much larger effect sizes on average than those without. If the same amount of data is collected from a study with overlap as a study without, the latter would have higher power. Moreover, as we saw in the main meta-analysis, certain constructions show much stronger effects than others. In particular, constructions where one form is very infrequent are likely to show large effects in logistic regression and produce low p-values. Therefore, there is a possibility that the estimated power in the p-curve analysis may be too high for many common study designs.

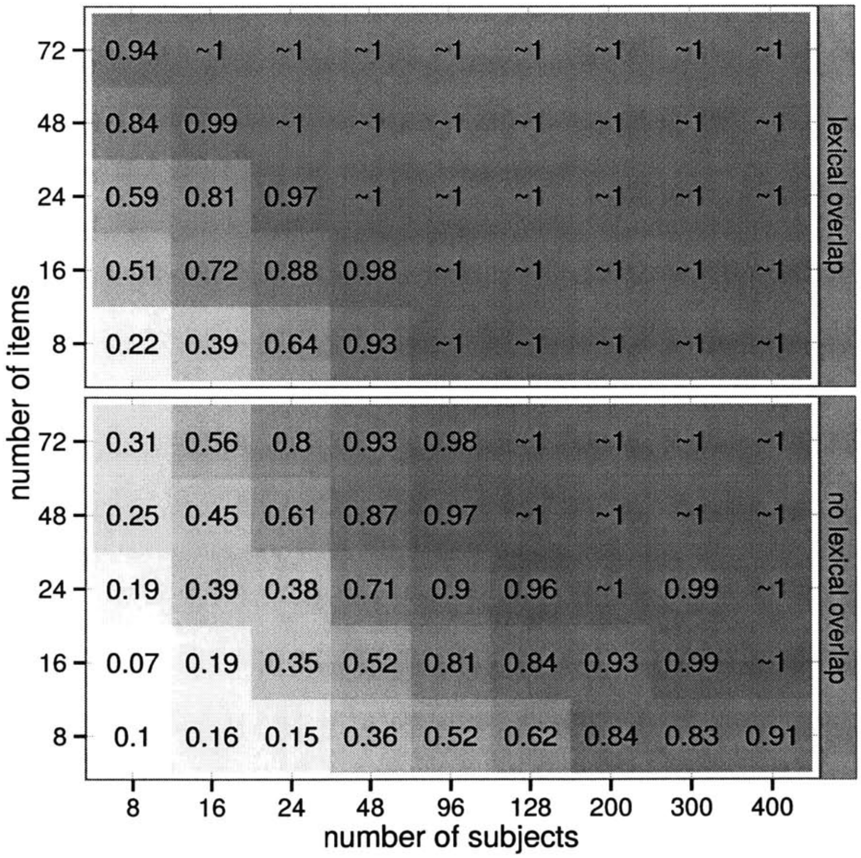
6.4 Sample size recommendations

Using the raw data collected from a subset of the papers in our sample, we can use simulation to give detailed recommendations for how to run future priming studies with sufficient statistical power. To do this, we used the mixed effect logistic regression described above in which we fit a regression to all data points from all studies for which we obtained raw data in order to simulate data for hypothetical new studies of varying designs. Specifically, we simulated 100 new experiments, each assumed to be a different random syntactic construction from a different experiment in a different paper. Each of the 100 experiments had S subjects and I items. The underlying "true" effect size was the effect size estimated using our actual data, and the effect size varied based on the paper, experiment, and subjects and items. We assume that 20% of data was lost due to "other" responses. We simulated experiments with all combinations of 8, 16, 24, 48, 96, 128, 200, 300, and 400 subjects and 8, 16, 24, 48, and 72 items.

Here, we assume that the underlying size of the "true" priming effect is .51 (change in log odds ratio) as estimated in the meta-analysis. However, using the raw data from the subset of 45 papers (or performing a meta-analysis on published results from those 45 papers), we find an effect size of only .34. Whether we assume the underlying true priming effect is .51 or .34 affects the power estimates. Here, we use the estimate of .51, which is based on more data. In the SI, we also provide sample size recommendations for when the underlying effect size is .51. For researchers performing syntactic priming studies, we recommend estimating the size of the expected effect based on the moderators (whether there is lexical boost, which construction is being used, etc.) and using an effect size appropriate to the task. The details of how we performed these simulations are in the SI.

First, we report the results of a simple two-cell design testing for the presence of syntactic priming in the absence of lexical overlap between prime and target. We show estimated statistical power in Figure 6-11. If power is 0, that means that when there is an underlying true effect of priming, we will detect it 0% of the time. If power is 1, that means that we will detect it 100% of the time. 80% is a standard threshold for power in experiments. To achieve that threshold comfortably for observing a simple priming effect, we recommend 96 subjects and 24 items (for 90% power).

Whereas many traditional power analyses in psychology focus on the number of subjects, psycholinguistics experiments typically have many items. As we show through these simulations, an increase in the number of items per participant substantially increases the power.



lexical overlap

number of items	8	16	24	48	96	128	200	300	400
72	0.94	~1	~1	~1	~1	~1	~1	~1	~1
48	0.84	0.99	~1	~1	~1	~1	~1	~1	~1
24	0.59	0.81	0.97	~1	~1	~1	~1	~1	~1
16	0.51	0.72	0.88	0.98	~1	~1	~1	~1	~1
8	0.22	0.39	0.64	0.93	~1	~1	~1	~1	~1

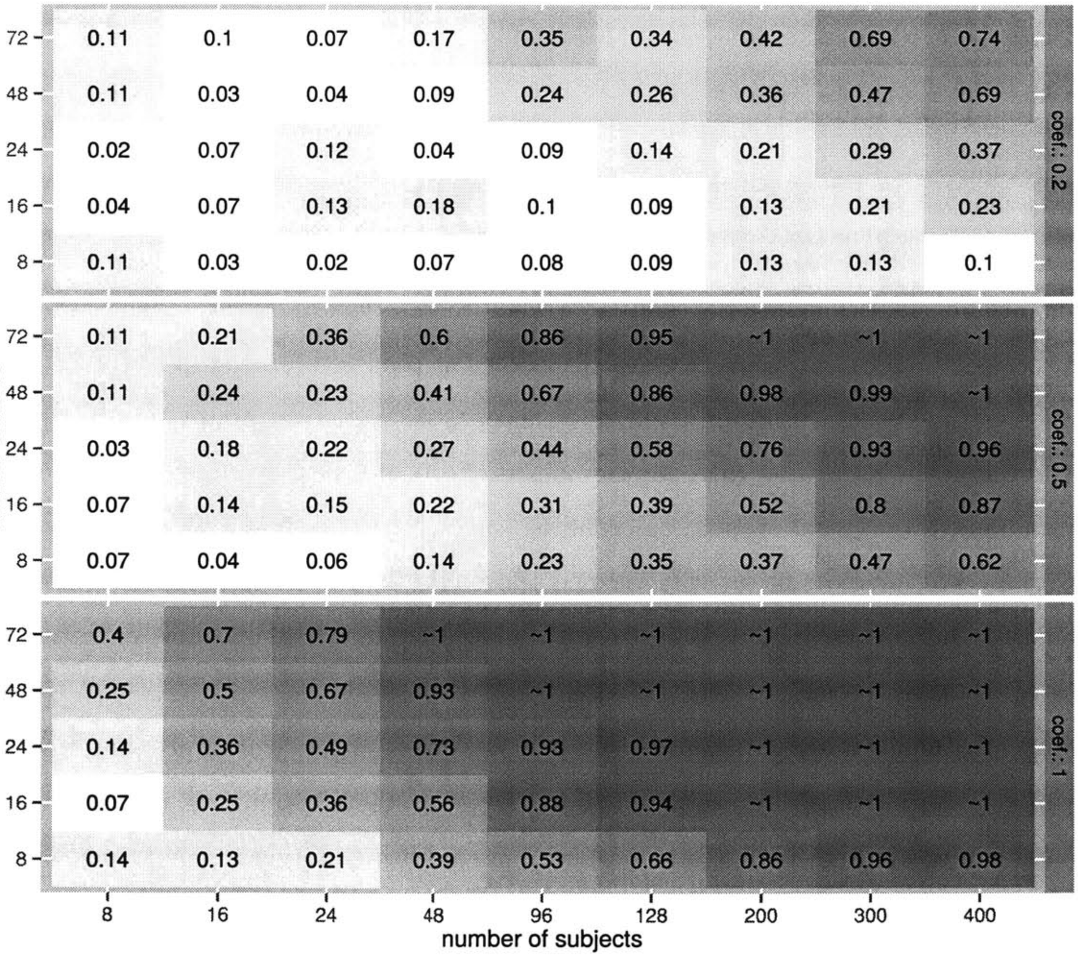
no lexical overlap

number of items	8	16	24	48	96	128	200	300	400
72	0.31	0.56	0.8	0.93	0.98	~1	~1	~1	~1
48	0.25	0.45	0.61	0.87	0.97	~1	~1	~1	~1
24	0.19	0.39	0.38	0.71	0.9	0.96	~1	0.99	~1
16	0.07	0.19	0.35	0.52	0.81	0.84	0.93	0.99	~1
8	0.1	0.16	0.15	0.36	0.52	0.62	0.84	0.83	0.91

Figure 6-11: Power to detect priming effect with lexical overlap (on top) and no lexical overlap (bottom).

When there is lexical overlap between the prime and target, the main effect of priming is much bigger and we need fewer subjects and items to have sufficient power, as shown in the bottom panel of Figure 6-11. Even with 16 subjects and 16 items, we have 92% power. Note that this number of subjects and items is not sufficient to detect a presence of a lexical overlap effect but merely to detect that priming exists when all the items repeat material between prime and target.

Next, we ask how many subjects and items we need to detect a moderator of priming (i.e., to detect an interaction), such as whether the priming effect is moderated by lexical overlap or lag. We try this for three different interaction sizes: coefficients of .2, .5, and 1. A coefficient of .2 is roughly the same order of magnitude as the interaction between prime and filler lag (a small but likely real interaction of theoretical interest). The coefficient of 1 is roughly the size of the lexical overlap effect. The coefficient .5 is somewhere in between. We show results for this analysis in Figure 6-12.



coef.	number of items	8	16	24	48	96	128	200	300	400
coef.: 0.2	72	0.11	0.1	0.07	0.17	0.35	0.34	0.42	0.69	0.74
	48	0.11	0.03	0.04	0.09	0.24	0.26	0.36	0.47	0.69
	24	0.02	0.07	0.12	0.04	0.09	0.14	0.21	0.29	0.37
	16	0.04	0.07	0.13	0.18	0.1	0.09	0.13	0.21	0.23
	8	0.11	0.03	0.02	0.07	0.08	0.09	0.13	0.13	0.1
coef.: 0.5	72	0.11	0.21	0.36	0.6	0.86	0.95	~1	~1	~1
	48	0.11	0.24	0.23	0.41	0.67	0.86	0.98	0.99	~1
	24	0.03	0.18	0.22	0.27	0.44	0.58	0.76	0.93	0.96
	16	0.07	0.14	0.15	0.22	0.31	0.39	0.52	0.8	0.87
	8	0.07	0.04	0.06	0.14	0.23	0.35	0.37	0.47	0.62
coef.: 1	72	0.4	0.7	0.79	~1	~1	~1	~1	~1	~1
	48	0.25	0.5	0.67	0.93	~1	~1	~1	~1	~1
	24	0.14	0.36	0.49	0.73	0.93	0.97	~1	~1	~1
	16	0.07	0.25	0.36	0.56	0.88	0.94	~1	~1	~1
	8	0.14	0.13	0.21	0.39	0.53	0.66	0.86	0.96	0.98

Figure 6-12: Power to detect an interaction with the priming effect, no lexical overlap.

Even with 400 subjects and 72 items, we do not achieve 80% power when the underlying interaction coefficient is .2. When it is .5 (a size larger than many of the interaction effects estimated in our meta-analysis), we need 96 subjects and 48 items to achieve over 80% power. And when the coefficient of interaction is large (1), we need fewer subjects.

Overall, these results suggest that, as our p-curve analysis also suggests, many of the studies in our sample are likely underpowered for the hypotheses that they are testing. Another important observation based on these simulations is that, even with a very large number of subjects, studies with fewer than 24 items are likely to be underpowered. Similarly, no matter how many items a study has, studies run with small numbers of participants (< 24) are likely to be underpowered.

6.5 Discussion

We conclude that there is strong evidence in the literature, over the last 30 years, for syntactic priming. We estimate that the size of this effect is small to medium when there is no lexical overlap and large when there is lexical overlap. The estimated effect is not likely the result of publication bias or p-hacking since most studies that investigate syntactic priming itself have acceptable statistical power. As has been reported in the literature, there are significant effects of lexical overlap, of lag between prime and target, and of bilingual priming–especially in its interaction with lexical overlap. We have quantitatively verified these effects and, harnessing the power of our large sample, estimated their size with more precision than any previous estimate provided.

While syntactic priming appears to be a robust effect, we found that the accumulated literature that studies moderators of syntactic priming suffers from low statistical power and, in the future, we recommend using larger sample sizes to study such phenomena. For that reason, we urge caution in interpreting studies that use only modestly sized samples to investigate whether some particular factor (other than lexical overlap, which leads to large effects) significantly affects the size of the syntactic priming effect. There have been cases in the literature where there are discrepancies whether an experiment finds significant effects of moderators like temporal lag or bilingualism. Given the typical sample sizes used in these experiments, it is possible that these discrepancies are mere noise and not reflective of any meaningful difference between the experiments.

It is also important to remember the limitations of a meta-analysis like this one. The results reported here are descriptive results of the syntactic priming literature sampled here. One might wonder, for instance, whether syntactic priming effects exist in the real world or only in laboratory settings. On that question, our meta-analysis has nothing to say. Nor can our meta-analysis answer whether the studies included here are in fact providing evidence for the various theories of syntactic priming and language processing more generally. There are also many variables, such as the types of fillers used, that were not included in our meta-analysis. We encourage future researchers to use the resources we have built here to ask and answer questions of their own devising. Our spreadsheets and materials used for these analyses are available on the Open Science Framework at `https://osf.io/b9zyk/`.

We hope that this work can be the basis of continuing meta-analysis and aggregation of data in syntactic priming. There is an extensive parallel literature on comprehension priming (where the dependent measure is sentence interpretation, reaction time, and so on) that could benefit from a similar sort of meta-analysis. Our meta-analysis also did not include unpublished work or work that appeared only at conferences. A meta-analysis of meta-analyses found systematic differences between unpublished work and peer-reviewed journal work such that the effect size of published studies is higher (Polanin, Tanner-Smith, & Hennessy, (2015), and it would be interesting to see whether the syntactic priming literature shows this effect. Our meta-analysis predicts that we should see only a moderate difference for studies investigating whether priming exists but perhaps a larger difference for studies investigating moderators of priming.

Chapter 7

Conclusion

7.1 On functional linguistics

The basic approach of the first part of this work was to investigate the large-scale structure of lexicons and to ask whether it fits with what we expect an efficient language to look like. To that end, the results in Chapters 2-4 are informative insofar as they explain the observed properties of languages–both within languages (e.g., word shortenings) and across languages (e.g., typological universals regarding phonotactics and frequency).

But what do we gain by understanding both how and why languages vary? To more fully appreciate the implications for cognition, it is worth pausing to investigate the assumptions of the approach used here. In effect, we have adopted a functionalist approach throughout this work, using the structure of language to illuminate the constraints on the human language processing mechanism. In that respect, this work joins a large body of literature in the functionalist tradition that uses the structure of language as a way of gaining insight into language processing and cognition more generally (Jespersen, 1922; Zipf, 1949; Hockett, 1960; Hawkins, 1994; Aylett & Turk, 2004; Givón, 2009; T. Jaeger & Tily, 2010; Kirby et al., 2015).

The general approach involves identifying features of language, both within and across languages, and trying to understand the functions of those features. As an example, consider the case of language universals: features that are universal across language. While there have been many language universals proposed (Chomsky, 1957; Greenberg, 1963; Dryer, 1992), human languages can vary enormously in virtually every dimension (Evans & Levinson, 2009). There are languages like English in which the order of words is largely fixed, and there are languages like Warlpiri in which the word order is quite free. Some languages have enormous inventories of distinct sounds, some languages have just a handful. But all human languages have something in common: every naturally occurring human language can be learned by babies in a matter of years and eventually used to communicate abstract thoughts and feelings with high fidelity. By understanding the properties of language that allow this to happen (properties that must be shared cross-linguistically), we can understand something about how the mind works.

While the statistical approaches used here are often thought of as distinct from mainstream linguistics, the end aims align with many of the aims of generative linguistics. Jerry Fodor summarized the goal of 20th century Chomskyan linguistics as being to "subtract the ways that human languages can differ from the ways in which it is conceivable that languages could differ" and to "subtract the information that is in the environment from the information that is required for the child to achieve linguistic mastery" (Fodor, 2001). The hope was that what would remain after these two subtractions is roughly equal and would represent the language endowment common to all humans. Fodor asks, "So why, you might wonder, didn't somebody just get a grant and do it?"

There are a few answers to Fodor's question. First, it's a hard problem that requires huge amounts of data. While the problem of natural language has still has not been reduced to one grant's worth of work, in recent years there has been progress made in developing a rigorous, quantitative understanding of how languages do and do not vary. The source of this progress has come, in part, from massive increases in the availability of structured linguistic data. But we cannot assume that, just because a property of language is universal to all languages, it must be part of our genetic endowment. Rather, it could be the case that a property is common to all languages because it allows the language to be efficient, to convey information rapidly while still being robust to noise in the perceptual system or in the environment.

A natural research program follows from these goals: first, to identify properties within and across languages; second, to ask whether those similarities are best explained by communicative factors; and third, to investigate what features of language are a result of our shared cognition. By identifying and separating those factors, we can being to flesh out a picture of the interaction among human cognition, communicative pressures, and the observed structure of language.

In the first part of this thesis, I focused on trying to make progress towards these goals in the restricted domain of the lexicon. Descriptively, I showed that words shorten after predictive contexts and that almost all world languages show a relationship between phonotactic probability and frequency. Using ideas from information theory, I showed that communicative factors can explain the way that words shorten in English, the cross-linguistic tendency for phonotactically words to be probable. And, using simulated lexicons from a null statistical model, I presented results that suggest that human cognition may well favor word forms that are clustered in phonetic space (although more work is needed to assess this possibility) and that the relationship between word length and frequency forms a part of speakers' knowledge of language.

As a result, this work can shed light on several existing claims about human language processing. First, it lends support to the now accepted idea that listeners are constantly updating predictions as to what word will come next (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Hale, 2001; Kamide, Altmann, & Haywood, 2003; Levy, 2008b; Demberg, 2010; N. J. Smith & Levy, 2013; Piantadosi, Tily, & Gibson, 2009; Tily & Piantadosi, 2009) and moreover that these predictions can drive language change. We saw this in the limited domain of word shortenings, but it is likely that predictability underlies many other sorts of language change from word order to syntax (Fedzechkina et al., 2012; Futrell, Mahowald, & Gibson, 2015).

Second, the widespread "clumpiness" of language suggests a plausible cognitive advantage for speakers to learn words that are phonetically similar to what they already know. Indeed, in other work, we show that this effect is even greater when the words are also semantically similar (Dautriche, Mahowald, Piantadosi, & Gibson, 2016, under revision) (see also Monaghan, Christiansen, and Chater (2007); Monaghan et al. (2011); Gasser (2004)) and that speakers have an easier time associating novel meanings with existing words than learning entirely novel words (Mahowald, Gibson, Alper, & Piantadosi, 2015). Taken together, these results argue against a strictly noisy channel account of lexical structure in which words are kept maximally distinct in phonetic space. In future experiments, we hope to test more directly the mechanisms by which phonetic and semantic material are stored together in memory – and whether shared storage could explain the preference for a phonetically clustered lexicon.

7.2 On functional scientific methods

Another answer to Fodor's question as to why somebody hasn't just "solved language" is that the problem has posed more methodological challenges than originally expected. Indeed, while it is relatively easy to access linguistic data, both traditional linguistic methods and experimental psychology methods have faced significant hurdles. With major questions about the reliability of data in linguistics (Gibson & Fedorenko, 2010b) and in the experimental sciences(Open Science Collaboration, 2015; Gelman & Loken, 2013; Ioannidis et al., 2014), it has become essential that we not just pursue new science but that we also carefully interrogate our methods. To that end, I chose to include two large-scale meta-investigations of methods: one in theoretical linguistics, one in experimental psychology. Both uncover serious issues in their respective fields and make concrete, actionable recommendations to researchers.

Ultimately, the methodological goals of the two fields is to obtain replicable results. A linguistic acceptability judgment should be robust across multiple speakers and dialects. A syntactic priming experiment that is studying a real effect should give the same results across speakers, labs, and paradigms. In linguistics, we now know that many acceptability judgments are not in fact reliable or robust. And, in psychology, many experiments do not replicate.

The question then becomes how we can we make sure that we are obtaining reliable results. A consistent theme that emerges in our meta-science is the need for baselines. In Chapter 4, we asked what a baseline *lexicon* looks like and whether real lexicons differ from these baselines. In Chapter 5, we asked what baseline linguistic judgments look like and used them to make recommendations about the collection of future judgments. While many judgments are highly reliable, there are a handful of erroneous judgments that pollute the literature. Without formal methods, it is hard to know which ones they are. In Chapter 6, we developed quantitative models about what the baseline syntactic priming study looks like–and what study design features are likely to cause deviations from that baseline.

By developing these methodological baselines, we can more easily integrate new studies into the existing bedrock of literature in linguistics and psychology. It is particularly illuminating to compare the work in Chapter 5 evaluating linguistic judgment paradigms to the work in Chapter 6 evaluating syntactic priming. The contrast in the methodologies of the two fields–qualitative linguistic judgments in linguistics and formal experimental methods in psychology–has been much debated in the linguistics community. Specifically, it has been claimed that linguistics avoids the problem of bad data by avoiding formal empirical data altogether and thus avoiding p-hacking and other questionable statistical practices (Hornstein, 2015).

It is true that some meta-analyses in psychology have uncovered large swaths of the psychology literature are not reproducible. But the meta-analysis in Chapter 6 (which found that syntactic priming was largely a robust effect) exemplifies one of the benefits of the experimental method: the existence of large bodies of empirical data means the ability to synthesize findings over many years. Such a meta-analysis would not be possible using traditional acceptability judgments:"informal" data cannot be easily meta-analyzed.

On the other hand, formal data collected for syntactic priming studies that is *not* publicly accessible–or at least not well enough described in the published report that it can be satisfactorily reconstructed–is also not sufficient for meta-analysis. To that end, we advocate that linguists, psycholinguists, and psychologists continue moving towards open data and open analysis scripts. By doing so, we can more readily avoid the pitfalls described in Part II of this thesis and allow the self-correcting nature of science to flourish.

References

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(01), 9–41.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*.

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010, March). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, *20*(03), 679–685. Retrieved from `http://www.worldscientific.com/doi/abs/10.1142/S021812741002596X` doi: 10.1142/S021812741002596X

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, *23*(4), 495–527.

Arppe, A., & Järvikivi, J. (2007). Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory*, *3*(1), 99–109.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048–3058.

Baayen, H. (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics* (pp. 271–278).

Baayen, R. (2001). *Word frequency distributions* (Vol. 1). Kluwer Academic Publishers.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using r.* Cambridge Univ Pr.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].*

Baayen, R., Piepenbrock, R., & van H, R. (1993). The CELEX lexical data base on cd-rom. *n.s.*

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

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., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*, 1001.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311. Retrieved from http://muse.jhu.edu/content/crossref/journals/language/v080/80.2bergen.pdf doi: 10.1353/lan.2004.0056

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2007). Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 931. Retrieved from http://psycnet.apa.org/journals/xlm/33/5/931/

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2012). Effects of phonological feedback on the selection of syntax: Evidence from between-language syntactic priming. *Bilingualism: Language and Cognition*, *15*(03), 503–516. Retrieved from http://journals.cambridge.org/abstract_S1366728911000162

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, *127*(3), 287–306. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0010027713000334 doi: 10.1016/j.cognition.2013.02.005

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence* (Vol. 25). Springer Science & Business Media.

Biria, R., Ameri-Golestan, A., & Antón-Méndez, I. (2010). Syntactic priming effects between modalities: A study of indirect questions/requests among persian english learners. *English Language Teaching*, *3*(3), p111. Retrieved from http://www.ccsenet.org/journal/index.php/elt/article/view/7221

Bloomfield, L. (1933). *Language.* New York: Henry Holt.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/0010028586900046 doi: 10.1016/0010-0285(86)90004-6

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Wiley.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. doi: 10.1016/j.jecp.2012.01.005

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. (2007, August). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163–197. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0010027706001247` doi: 10.1016/j.cognition.2006.05.006

Branigan, H. P., Pickering, M. J., Stewart, A. J., & McLean, J. F. (2000). Syntactic priming in spoken production: Linguistic and temporal interference. *Memory & Cognition*, *28*(8), 1297–1302. Retrieved from `http://link.springer.com/article/10.3758/BF03211830`

Bunger, A., Papafragou, A., & Trueswell, J. C. (2013). Event structure influences language production: Evidence from structural priming in motion event description. *Journal of Memory and Language*, *69*(3), 299–323. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X13000259` doi: 10.1016/j.jml.2013.04.002

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013a). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*, *14*(8), 585–585.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013b). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in french 18-month-olds. *Language Learning and Development*, *10*(1), 1–18.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1975). *Reflections on language.* Pantheon Books.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* Praeger Publishers.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*, 97-138.

Cleland, A. (2003, August). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214–230. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X03000603` doi: 10.1016/S0749-596X(03)00060-3

Cleland, A., & Pickering, M. J. (2006, February). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, *54*(2), 185–198. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X05001269` doi: 10.1016/j.jml.2005.10.003

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213. doi: 10.1016/j.jecp.2004.07.004

Cohen, J. (1977). *Statistical power for the behavioral sciences.* Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Cohen Priva, U. (2008). Using Information Content to Predict Phone Deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics* (p. 90).

Conrad, B., & Mitzenmacher, M. (2004). Power laws for monkeys typing randomly: the case of unequal probabilities. *Information Theory, IEEE Transactions on*, *50*(7), 1403–1414.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments.* Sage publications.

Coyle, J. M., & Kaschak, M. P. (2012). Female fertility affects men's linguistic choices. *PLoS ONE*, *7*(2), e27971. Retrieved from `http://dx.plos.org/10.1371/journal.pone.0027971` doi: 10.1371/journal.pone.0027971

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.

Culicover, P. W., & Jackendoff, R. (2010, June). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, *14*(6), 234–235. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S1364661310000707` doi: 10.1016/j.tics.2010.03.012

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (submitted). Wordform similarity increases with semantic similarity: an analysis of 101 languages. *Cognitive Science*.

Dautriche, I., Mahowald, K., Piantadosi, S. T., & Gibson, E. (2016, under revision). Word forms are structured for efficient use. *Cognitive Science*.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, *6*, 9.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 611–629.

Demberg, V. (2010). *A Broad-Coverage Model of Prediction in Human Sentence Processing* (Unpublished doctoral dissertation). University of Edinburgh.

Dryer, M. (1992). The Greenbergian word order correlations. *Language*, *68*(1), 81–138.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Evans, N., & Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429–448.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012, October). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1215776109 doi: 10.1073/pnas.1215776109

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Ferreira, V., Bock, K., Wilson, M. P., & Cohen, N. J. (2008). Memory for syntax despite amnesia. *Psychological Science*, *19*(9), 940–946. Retrieved from http://pss.sagepub.com/content/19/9/940.short

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and cognitive processes*, *21*(7-8), 1011–1029.

Ferrer i Cancho, R., & Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, *2011*, L12002.

Ferrer-i Cancho, R., & Moscoso del Prado Martín, F. (2011, December). 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. *Journal of Statistical Mechanics: Theory and Experiment*, *41*(4), L12002. Retrieved from http://www.springerlink.com/index/10.3758/BRM.41.4.977 doi: 10.3758/BRM.41.4.977

Flemming, E. (2002). *Auditory Representations in Phonology.* Routledge.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In *Phonetically Based Phonology.* Cambridge: Cambridge University Press.

Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology.* MIT press.

Francis, G., Tanzman, J., & Matthews, W. J. (2014, December). Excess Success for Psychology Articles in the Journal Science. *PLoS ONE*, *9*(12), e114255. Retrieved from http://dx.plos.org/10.1371/journal.pone.0114255 doi: 10.1371/journal.pone.0114255

Frank, A., & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Cogsci. Washington, DC: CogSci.*

Frauenfelder, U., Baayen, R., & Hellwig, F. (1993, December). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*(6), 781–804. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X83710399 doi: 10.1006/jmla.1993.1039

Futrell, R., Mahowald, K., & Gibson, E. (2015). Quantifying word order freedom in dependency corpora. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 91–100.

Gafos, A. I. (2014). *The articulatory basis of locality in phonology.* Routledge.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the easy/hard database. *Journal of Phonetics*, *49*, 96–116.

Gahl, S., & Garnsey, S. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, *80*(4), 748–775.

Gahl, S., Yao, Y., & Johnson, K. (2012, May). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and*

Language, *66*(4), 789–806. doi: 10.1016/j.jml.2011.11.006

Gasser, M. (2004). The origins of arbitrariness in language. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 26, pp. 4–7).

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition'or 'p-hacking'and the research hypothesis was posited ahead of time. *Downloaded January*, *30*, 2014.

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of empirical methods in natural language processing* (pp. 65–72).

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences.* Retrieved from `http://www.pnas.org/content/early/2013/05/01/1216438110.abstract` doi: 10.1073/pnas.1216438110

Gibson, E., & Fedorenko, E. (2010a). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes.*

Gibson, E., & Fedorenko, E. (2010b, June). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*(6), 233–234. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S1364661310000525` doi: 10.1016/j.tics.2010.03.005

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1-2), 88–124.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*(8), 509–524.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229–240.

Givón, T. (2009). *The Genesis of Syntactic Complexity: Diachrony, Ontogeny, Neuro-cognition, Evolution.* John Benjamins Publishing Co.

Goldrick, M., & Rapp, B. (2002). A restricted interaction account (ria) of spoken word production: The best of both worlds. *Aphasiology*, *16*(1-2), 20–55.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, *30*(3), 859–896. doi: 10.1007/s11049-012-9169-1

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289. Retrieved from `http://doi.wiley.com/10.1111/j.1756-8765.2012.01186.x` doi: 10.1111/j.1756-8765.2012.01186.x

Graff, P. (2012). *Communicative efficiency in the lexicon* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. , 73–113.

Gross, S., & Culbertson, J. (2011). Revisited linguistic intuitions. *The British Journal for the Philosophy of Science*, *62*(3), 639–656.

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).

Hartsuiker, R. J. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology: Section A*, *52*(1), 129–147. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/713755798

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214–238. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X07000861 doi: 10.1016/j.jml.2007.07.003

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, *15*(6), 409–414. Retrieved from http://pss.sagepub.com/content/15/6/409.short

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*(2), B27–B39. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027799000803

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167.

Hawkins, J. (1994). *A performance theory of order and constituency* (Vol. 73). Cambridge, UK: Cambridge University Press.

Hayes, B. (2012). *BLICK - a Phonotactic Probability Calculator.*

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. doi: 10.1162/ling.2008.39.3.379

Heydel, M., & Murray, W. S. (2000). Conceptual effects in sentence priming: A cross-linguistic perspective. In *Cross-linguistic perspectives on language processing* (pp. 227–254). Springer.

Hinton, L., Nichols, J., & Ohala, J. J. (2006). *Sound Symbolism.* Cambridge University Press.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Hockett, C., & Voegelin, C. (1955). *A manual of phonology* (Vol. 21) (No. 4). Waverly Press Baltimore, MD.

Hornstein, N. (2015). *Bad Data.* Retrieved from facultyoflanguage.blogspot.com/2013/01/bad-data.html

Householder, F. W. (1965). On some recent claims in phonological theory. *Journal of Linguistics*, *1*(01), 13–34.

Howes, D. (1968). Zipf's Law and Miller's Random-Monkey Model. *The American Journal of Psychology*, *81*(2), 269–272.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298–20130298. doi: 10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65. doi: 10.1016/j.cognition.2008.07.015

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, *18*(5), 235–241.

Jaeger, T. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech* (Unpublished doctoral dissertation). Stanford University.

Jaeger, T. (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. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T., & Tily, H. (2010). On language 'utility': processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Jaeger, T. F., & Snider, N. E. (2013, April). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0010027712002636` doi: 10.1016/j.cognition.2012.10.013

Jespersen, O. (1922). *Language: Its nature, development, and origin.* New York: Henry Holt and Co.

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proc. of NAACL-HLT* (pp. 139–146).

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency in the emergence of linguistic structure.* Amsterdam: John Benjamins.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, N.J.: Pearson Prentice Hall.

Jusczyk, P., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645. doi: 10.1006/jmla.1994.1030

Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kantola, L., & van Gompel, R. P. G. (2011). Between- and within-language priming is the same: Evidence for shared bilingual syntactic representations. *Memory & Cognition*, *39*(2), 276–290. Retrieved from `http://link.springer.com/10.3758/s13421-010-0016-5` doi: 10.3758/s13421-010-0016-5

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition*, *35*(5), 925–937. Retrieved from http://link.springer.com/article/10.3758/BF03193466

Kaschak, M. P., & Borreggine, K. L. (2008, April). Is long-term structural priming affected by patterns of experience with individual verbs? *Journal of Memory and Language*, *58*(3), 862–878. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X07000137 doi: 10.1016/j.jml.2006.12.002

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011, December). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, *18*(6), 1133–1139. Retrieved from http://www.springerlink.com/index/10.3758/s13423-011-0157-y doi: 10.3758/s13423-011-0157-y

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition*, *99*(3), B73–B82. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0010027705001241 doi: 10.1016/j.cognition.2005.07.002

Kawasaki, H., & Ohala, J. J. (1980). Acoustic basis for universal constraints on sound sequences. *The Journal of the Acoustical Society of America*, *68*(S1), S33–S33.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.99.2.349 doi: 10.1037/0033-295X.99.2.349

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.

Kim, Y., & McDonough, K. (2007). Learners' production of passives during syntactic priming activities. *Applied Linguistics*, *29*(1), 149–154. Retrieved from http://applij.oxfordjournals.org/cgi/doi/10.1093/applin/amn004 doi: 10.1093/applin/amn004

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. Retrieved from http://www.pnas.org/content/105/31/10681.short

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kootstra, G. J., van Hell, J. G., & Dijkstra, T. (2010, August). Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language*, *63*(2), 210–231. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X10000288 doi: 10.1016/j.jml.2010.03.006

Labov, W. (1978). Sociolinguistics. *A Survey of Linguistic Science*.

Landauer, T., & Streeter, L. (1973, April). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119–131. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0022537173800015` doi: 10.1016/S0022-5371(73)80001-5

Landy, J., & Goodwin, G. (2014). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Manuscript submitted for publication.*

Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychological Science*, *17*(4), 273–277.

Levelt, W. J., Van Gent, J., Haans, A., & Meijers, A. (1977). Grammaticality, paraphrase, and imagery. *Acceptability in language*, 87–101.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady* (Vol. 10, p. 707).

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 conference on empirical methods in natural language processing* (pp. 234–243).

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, *19*, 849–856.

Lindblom, B. (1983). *Economy of Speech Gestures.* Springer.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech Production and Speech Modelling* (pp. 403–439). Springer.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. *Phonological development: Models, research, implications*, 131–163.

Linzen, T., & Oseki, Y. (2015). *The reliability of acceptability judgments across languages.*

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis* (Vol. 49). Sage Publications Thousand Oaks, CA.

Luce, P. A. (1986). *Neighborhoods of Words in the Mental Lexicon* (Unpublished doctoral dissertation). Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1), 1.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156. doi: 10.1080/03640210709336987

Mahowald, K., Dautriche, I., Piantadosi, S. T., Christophe, A., & Gibson, E. (Under revision). Lexical clustering in efficient language design.

Mahowald, K., Dautriche, I., Piantadosi, S. T., & Gibson, E. (Under revision). Word forms are structured for efficient use. *Cognitive Science*.

Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2012, October). Info/information theory: Speakers choose shorter words in predictive contexts.

Cognition. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0010027712002107 doi: 10.1016/j.cognition.2012.09.010

Mahowald, K., Gibson, E., Alper, M., & Piantadosi, S. T. (2015). *Lexical items are privileged slots for meaning.* (Poster presented at 2Annual CUNY Conference on Human Sentence Processing, CUNY 2015)

Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2015). SNAP Judgments: A Small N Acceptability Paradigm (SNAP) for Linguistic Acceptability Judgments. *Language*.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in production. *Journal of Memory and Language*.

Mandelbrot, B. (n.d.). An informational theory of the statistical structure of language. *Communication theory*, 486–502.

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *J. Information Processes*, *6*, 229–236.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.

Marchand, H. (1966). *The categories and types of present-day english word-formation: A synchronic-diachronic approach* (No. 13). University of Alabama Press.

Mason, W., & Suri, S. (2012, March). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. Retrieved from http://www.springerlink.com/index/10.3758/s13428-011-0124-6 doi: 10.3758/s13428-011-0124-6

McCawley, J. D. (1982). How far can you trust a linguist. *Language, Mind, and Brain*, 75–87.

Messenger, K. (2010). Syntactic priming and children's production and representation of the passive. *Language Acquistion*, *17*, 121-123.

Miller, G. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 311–314.

Miller, G., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0022924 doi: 10.1037/a0022924

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*.

Myers, J. (2009, March). The design and analysis of small-scale syntactic judgment experiments. *Lingua*, *119*(3), 425–444. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0024384108001617 doi: 10.1016/j.lingua.2008.09.003

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, *4*(02), 115–125. doi: 10.1515/langcog-2012-0007

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*(1), 181–186. doi: 10.1016/j.cognition.2009.04.001

Open Science Collaboration. (2015, August). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. Retrieved from `http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716` doi: 10.1126/science.aac4716

Pappert, S., & Pechmann, T. (2013). Bidirectional structural priming across alternations: Evidence from the generation of dative and benefactive alternation structures in german. *Language and Cognitive Processes*, *28*(9), 1303–1322. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/01690965.2012.672752` doi: 10.1080/01690965.2012.672752

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17.

Piantadosi, S. (2014, October). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. doi: 10.3758/s13423-014-0585-6

Piantadosi, S., Tily, H., & Gibson, E. (2009). The communicative lexicon hypothesis. In *The 31st annual meeting of the cognitive science society (CogSci09)* (p. 2582–2587).

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S., Tily, H., & Gibson, E. (2012, March). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0010027711002496` doi: 10.1016/j.cognition.2011.10.004

Piantadosi, S., Tily, H., & Gibson, E. (2013). Information content versus word length in natural language: A reply to ferrer-i-cancho and moscoso del prado martin [arXiv: 1209.1751]. *arXiv preprint arXiv:1307.6726*.

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. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X9892592X` doi: 10.1006/jmla.1998.2592

Pickering, M. J., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459. Retrieved from http://psycnet.apa.org/journals/bul/134/3/427.html doi: 10.1037/0033-2909.134.3.427

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2015). Estimating the difference between published and unpublished effect sizes a meta-review. *Review of Educational Research*, 0034654315582067.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265–282. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X97925468

R Core Team. (n.d.). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raymond, W. D., Dautricourt, R., & Hume, E. (2006). Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, *18*(01), 55–97.

Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, *68*, 33–58.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027703001197

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language*, *56*(2), 153–171. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X06001471

Schutze, C. (2006). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* Chicago: University of Chicago Press.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*(2), 228–264.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 623–656.

Shin, J.-A., & Christianson, K. (2012). Structural priming and second language learning: Structural priming and l2 learning. *Language Learning*, *62*(3), 931–964. Retrieved from http://doi.wiley.com/10.1111/j.1467-9922.2011.00657.x doi: 10.1111/j.1467-9922.2011.00657.x

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-Curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Slocombe, K. E., Alvarez, I., Branigan, H. P., Jellema, T., Burnett, H. G., Fischer, A., ... Levita, L. (2013, June). Linguistic Alignment in Adults with and Without Asperger's Syndrome. *Journal of Autism and Developmental Disorders*, *43*(6), 1423–1436. Retrieved from http://link.springer.com/10.1007/s10803-012-1698-2 doi: 10.1007/s10803-012-1698-2

Smith, K., Kirby, S., & Brighton, H. (2003, October). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/106454603322694825 doi: 10.1162/106454603322694825

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sorace, A., & Keller, F. (2005, November). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0024384104001032 doi: 10.1016/j.lingua.2004.07.002

Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research*, *2*(2), 83–98. Retrieved from http://link.springer.com/10.1007/BF01067203 doi: 10.1007/BF01067203

Sporns, O. (2011). *Networks of the Brain.* MIT Press.

Sprouse, J. (2011, March). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167. Retrieved from http://www.springerlink.com/index/10.3758/s13428-010-0039-7 doi: 10.3758/s13428-010-0039-7

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1), 413–422.

Steriade, D. (1997). *Phonetics in phonology: The case of laryngeal neutralization.*

Steriade, D. (2001). Directional asymmetries in place assimilation: A perceptual account. In *In hume and johnson.*

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(02). doi: 10.1017/S0142716404001109

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(02), 291. doi: 10.1017/S030500090800891X

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192.

Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken american english. *Behavior Research Methods*, *42*(2), 497–506. doi: 10.3758/BRM.42.2.497

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of child language*, *32*(4), 827.

Storkel, H. L., Maekawa, J., & Aschenbrenner, A. J. (2012). The Effect of Homonymy on Learning Correctly Articulated Versus Misarticulated Words. *Journal of Speech, Language, and Hearing Research*, *56*(2), 694–707.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99. Retrieved from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613642/`

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference. Amsterdam, The Netherlands.*

Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*.

Van Son, R., & Pols, L. (2003). How efficient is speech? *Proc. of the Institute of Phonetic Sciences*, *25*, 171–184.

van Gompel, R. P., Arai, M., & Pearson, J. (2012, February). The representation of mono- and intransitive structures. *Journal of Memory and Language*, *66*(2), 384–406. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0749596X11001173` doi: 10.1016/j.jml.2011.11.005

van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings Institute of Phonetic Sciences, University of Amsterdam*, *25*, 171–184.

Van Son, R. J., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*(1), 100–123.

Verreyt, N., Bogaerts, L., Cop, U., Bernolet, S., De Letter, M., Hemelsoet, D., ... Duyck, W. (2013, July). Syntactic priming in bilingual patients with parallel and differential aphasia. *Aphasiology*, *27*(7), 867–887. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/02687038.2013.791918` doi: 10.1080/02687038.2013.791918

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from `http://www.jstatsoft.org/v36/i03/`

Vitevitch, M. S. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1-2), 306–311. doi: 10.1006/brln.1999

.2116

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, *67*(1), 30–44.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, *31*(4), 491–504.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*(11), 1481–1496.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n3. *Information and control*, *10*(2), 189–208.

Zipf, G. (1935). The Psychology of Language. *NY Houghton-Mifflin*.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* New York: Addison-Wesley.