UNIVERSITY OF CALIFORNIA SAN DIEGO

# Ambiguity in the Mind and the Lexicon

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Cognitive Science

by

Sean Trott

Committee in charge:

Professor Benjamin Bergen, Chair
Professor Seana Coulson
Professor Victor Ferreira
Professor Marta Kutas
Professor Ndapa Nakashole

2022

The Dissertation of Sean Trott is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

# DEDICATION

For my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# ACKNOWLEDGEMENTS

The work in this thesis could not have been completed without the help and support of others.

First, thank you to my advisor, Benjamin Bergen, whose skill is unparalleled when it comes to taking vague or half-formed ideas and identifying a practicable kernel of insight—something tractable and testable that touches empirical data in some way. Ben's first question is almost always *why*—why ask this question rather than that one—and it's a question I aim to keep asking. He's taught me the importance of sketching out the structure of an argument, even or especially if it's one I disagree with. I'd also like to thank him for his patience and kindness through the years.

Thank you also to the rest of my committee (Vic Ferreira, Marta Kutas, Seana Coulson, and Ndapa Nakashole) and to other collaborators here at UCSD (Federico Rossano, Katherine DeLong, Arturs Semenuks, James Michaelov, Cameron Jones, Tyler Chang, Stefanie Reed, Alex Liebscher). I've learned so much from all of you.

Thank you to my cohort, to the members of the Language and Cognition lab, and to the other friends I've been fortunate enough to make at UCSD. I will always treasure our many conversations in the office and beyond, which have challenged and shaped my own worldview and were also just a lot of fun. One of the things I'll miss most about my time here is those unexpected afternoon encounters that somehow spill into two-hour chats at the coffee cart. Thoughts mean more when you can think them with others.

Thank you to my family: you made me feel capable of pursuing the things I wanted to pursue, and I'll always be grateful to that. Thank you to friends—especially Josh Duffy—for your support.

Thank you most of all to Pamela Rivière. There's much I could say but language, as is so frequently the case, just doesn't feel up to the task. So I'll just say this: even if none of the above were true, coming to UCSD would always have been worth it because it's how I met you.

Chapter 2, in full, is a reprint of the material as it appears in the Proceedings of the 59th Annual Meeting of the Association for computational Linguistics and the 11th International Joint Conference on Natural Language Processing, in 2021 (August). Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted to the 29th International Conference on Computational Linguistics. Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is under review at *Psychological Review*. Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in *Cognition*. Trott, Sean; Bergen, Benjamin (2020). The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it will appear in the Proceedings of the 44th Annual Conference in Cognitive Science. Trott, Sean; Bergen, Benjamin (2022). The dissertation author was the primary investigator and author of this paper.

Chapter 7, in full, is a reprint of the material as it appears in *Cognition*. Trott, Sean; Bergen, Benjamin (2022). The dissertation author was the primary investigator and author of this paper.

# VITA

2014 Bachelor of Arts in Cognitive Science, UC Berkeley

2022 Doctor of Philosophy in Cognitive Science, University of California San Diego

ABSTRACT OF THE DISSERTATION

Ambiguity in the Mind and the Lexicon

by

Sean Trott

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2022

Professor Benjamin Bergen, Chair

Words contain multitudes. This multiplicity of meanings raises two key questions, both of which this thesis attempts to address. First, are word meanings categorical or continuous? The results of Chapters 2-4 support a hybrid model, in which word meanings occupy a continuous state-space (Elman, 2009), which is further discretized along the

boundaries of distinct senses. And second, does the amount and distribution of homophony in real lexica reflect a pressure to concentrate meanings in the most efficient, optimal wordforms? The results in Chapters 5-7 suggest that homophony can emerge without a direct pressure for efficiency—and further, that real lexica might select *against* homophones, particularly among the most frequent wordforms of a lexicon. This pressure could even explain other properties of human lexica, such as their large phonological neighborhoods.

# CHAPTER 1: INTRODUCTION

Words mean different things in different contexts. In some cases (approximately 7% of English words, for instance—Rodd et al., 2004), these meanings appear entirely unrelated: the "bark" of a dog is utterly unlike the "bark" of a tree. Far more frequently (about 84% of English words—Rodd et al., 2004), the same wordform conveys distinct but related meanings: a "lamb" can be friendly, but it can also be roasted. Finally, all words have meanings that arguably depend on context to some extent: a toddler and a cheetah can both "run", but the motor routines involved in each event—and a comprehender's representation of those events—likely differ in a number of ways (Elman, 2004; Yee & Thompson-Schill, 2016).

Words, then, contain multitudes. The resulting *lexical ambiguity* raises two related (but distinct) research questions, each of which connect to broader debates in the field of Cognitive Science. This thesis attempts to address both questions, which are introduced and described briefly in the sections below.[1]

**Are word meanings categorical or continuous?**

Despite widespread interest in lexical ambiguity across many domains—linguistics (Tuggy, 1993), cognitive science (Rodd et al., 2004), lexicography (Krishnamurthy & Nicholls, 2000), Natural Language Processing (Navigli, 2009; Kilgarriff, 2007; Schneider et al, 2015; Karidi et al, 2021), legal studies (Schane, 2002), and more—there remains considerable debate about exactly how humans represent this multiplicity of meanings.

[1] Note that while both questions relate to lexical ambiguity, the lines of research are fairly distinct in terms of the methodologies employed and the theoretical assumptions they make. This issue is discussed at more length in the Conclusion chapter.

Traditional views of the mental lexicon liken it to a kind of "mental dictionary", in which different word meanings are stored in discrete "entries" (Kempson, 1977). However, others (Elman, 2009) have criticized these categorical accounts on several grounds, including the fact that they are unable to handle the flexible, context-dependent nature of word meaning. Alternative accounts argue instead that word meanings are best characterized as occupying a continuous, context-sensitive state space (Elman, 2004). As noted above, this tension echoes more general debates in Cognitive Science. What is the nature of mental representations? Is semantic knowledge continuous or categorical? In chapters 2-4, I attempt to adjudicate between these categorical and continuous accounts of word meaning.

In Chapter 2, I introduce a novel dataset of human relatedness judgments for ambiguous words in distinct contexts. I also show that state-of-the-art language models trained on linguistic input alone make systematic errors in how related they find these meanings to be—they *underestimate* the relatedness of words belonging to the same sense, and *overestimate* the relatedness of different sense homonyms (Chapter 2). This suggests that at least as operationalized by current language models, continuous accounts fail to explain the categorical effect of sense boundaries.

In Chapter 3, I introduce an expanded version of this dataset that contains contextualized sensorimotor judgments about these ambiguous words (modeled on the Lancaster Sensorimotor Norms—Lynott et al., 2019). I show that these contextualized judgments encode novel information that is not present in either the Lancaster Norms or distributional language models (Chapter 3).

Finally, in Chapter 4, I directly test which account of word meaning best explains human behavior on a primed sensibility judgment task. The results from two behavioral experiments,

along with a quantitative model, suggest evidence for a "hybrid" theory, in which word meanings are *both* categorical and continuous (Chapter 4).

**Why is language so ambiguous?**

Ambiguity is surprisingly pervasive. Even considering homonyms alone, anywhere from approximately 7% (Rodd et al., 2015) to 15% (Trott & Bergen, 2020) of English wordforms have at least two unrelated meanings. This is surprising: why would a system ostensibly evolved for efficient communication tolerate such rampant ambiguity?

In Chapter 5, I consider a dominant theory that ambiguity actually makes languages *more* efficient, i.e., that ambiguity is positively selected for. Surprisingly, and contrary to previous work (Piantadosi et al, 2012), I find that an explanation for both the *amount* and *concentration* of homophony in real lexica need not posit a direct selection pressure for recycling the "best" wordforms for multiple meanings. Simulated lexica matched for the phonotactics and distribution of word lengths actually *overestimate* the degree of ambiguity—suggesting that, if anything, homophones are selected *against* (Chapter 5).

In Chapter 6, I ask whether this apparent pressure against homophony can explain other properties of real lexica, such as the size of their phonological neighborhoods, which previous work (Dautriche et al., 2017) has argued may also be the product of a direct selection pressure. Here, I find that implementing a selection pressure against homophones results in the creation of larger phonological neighborhoods, suggesting that at least in principle, both phenomena could be explained by a common mechanism (Chapter 6).

Finally, in Chapter 7, I consider the well-known *meaning-frequency law* (Zipf, 1949): the empirical observation that the most frequent wordforms are also the most ambiguous. Previous work (Zipf, 1949; Piantadosi et al., 2012) has argued that this law results from competing

pressures to make language easier to produce, on the one hand (i.e., a pressure for *unification*), and to make language easier to understand, on the other hand (i.e., a pressure for *diversification*). However, this work has not attempted to quantify the relative size of these pressures, leaving it unknown whether they are equal in magnitude or whether one is stronger than the other. Using a phonotactic baseline, I find evidence supporting the claim that the distribution of meanings across wordforms is shaped by a relatively stronger *comprehender-centric* pressure to ease the cost of frequent disambiguation (Chapter 7).

**Summary**

Multiplicity of meanings appears to be the rule, not an exception. The mere prevalence of this phenomenon demands that it be taken seriously—both for theories of how meanings are organized and represented in the mind, and for theories of why languages look the way that they do. The work in this thesis attempts to address this challenge by investigating two distinct questions about lexical ambiguity.

# CHAPTER 2: RELATEDNESS OF AMBIGUOUS WORDS—IN CONTEXT

Words mean different things in different contexts. Sometimes these meanings appear to be distinct, a phenomenon known as *lexical ambiguity*. In English, approximately 7% of wordforms are *homonymous*, i.e., they have multiple, unrelated meanings[2](e.g., "tree bark" vs. "dog bark"), and as many as 84% of wordforms are *polysemous*, i.e., they have multiple, related meanings (e.g., "pet chicken" vs. "roast chicken") (Rodd et al., 2004). But even unambiguous words evoke subtly different interpretations depending on the context of use, i.e., their meanings are dynamic and *context-dependent* (Yee and Thompson-Schill, 2016; Li and Joanisse, 2021). While the uses of *runs* in "the boy runs" vs. "the cheetah runs" may not be considered distinct meanings, a human comprehender will likely activate a different mental image when processing each sentence (Elman, 2009).

These facts present a challenge for computational models of lexical semantics. Any downstream task that involves meaning requires models capable of *disambiguating* among the multiple possible meanings of an ambiguous word in a given context. Further, the *graded* nature of human semantic representations can influence how comprehenders construe events and participants in those events (Elman, 2009; Li and Joanisse, 2021). In turn, a number of Natural Language Processing (NLP) tasks could benefit from context-sensitive representations that go beyond discrete sense representations and capture the manner in which humans construe events—including sentiment analysis, bias detection, machine translation, and more (Trott et al., 2020). If an eventual goal of NLP is human-like language understanding, models must be equipped with semantic representations that are flexible enough to accommodate the dynamic,

[2] Dautriche (2015) estimates the average rate of homonymy across languages to be 4%.

context-dependent nature of word meaning—as humans appear to do (Elman, 2009; Li and Joanisse, 2021).

Yet a crucial prerequisite to developing better models is *evaluating* those models along the relevant dimensions of performance. Thus, at the minimum, we need metrics that evaluate a model along two critical dimensions:

1. **Disambiguation**: A model's ability to distinguish between distinct meanings of a word.
2. **Contextual Gradation**: A model's ability to modulate a given meaning in context, in ways that reflect the continuous nature of human judgments.

A promising development in recent years is the rise of contextualized word embeddings, produced using neural language models such as BERT (Devlin et al., 2018), ELMo (Peters et al., 2018), XLNet (Yang et al., 2019), and more. Advances in these models have yielded improved performance on a number of tasks, including Word Sense Disambiguation (WSD) (Boleda et al., 2019; Loureiro et al., 2020).

WSD satisfies the Disambiguation Criterion above, but not the Contextual Gradation Criterion. In fact, there is still a dearth of metrics for assessing the degree to which contextualized representations match human judgments about the way in which context shapes meaning.

In the Related Work section, we describe several related datasets that satisfy at least one of these criteria. Then, we introduce and describe the dataset construction process for RAW-C: Relatedness of Ambiguous Words—in Context.[3] We also describe the procedure we followed for collecting human relatedness norms for each sentence pair. In the remaining sections, we report the results of several analyses that probe how well contextualized embeddings from two neural

[3] The dataset can be found on GitHub: https:// github.com/seantrott/raw-c.

language models (BERT and ELMo) predict these norms, then explore possible shortcomings in current models, and propose avenues for future work.

## Related Work

Most existing datasets fulfill either the Disambiguation or the Contextual Gradation criterion, but few datasets fulfill both (see Haber and Poesio (2020a) for an exception).

Several datasets contain human relatedness and similarity judgments for *distinct* words in isolation. Others are used for Word Sense Disambiguation, and contain ambiguous words in different sentence contexts, along with annotated sense labels; as noted in the Introduction, WSD fulfills the Disambiguation Criterion, but not the Contextual Gradation Criterion. Several recent datasets contain graded relatedness judgments for words in different contexts. However, none focus specifically on graded relatedness judgments for *ambiguous* words, controlling both the inflection and part of speech of the target word in question. Finally, one dataset (Haber and Poesio, 2020) contains similarity judgments for polysemous words in context, but is more limited in size and does not match the sentence frame across the two uses.

### De-contextualized Word Similarity and Relatedness

Several datasets contain human judgments of the *similarity* or *relatedness* of (mostly English) word pairs, in isolation (see Taieb et al. (2020) for a review). This includes SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016), WordSim-353 (Finkelstein et al., 2001), MTurk-771 (Halawi et al., 2012), MEN (Bruni et al., 2014), and more. These datasets are primarily used for evaluating the quality of static semantic representations, including distributed semantic models such as GloVe (Pennington et al., 2014), as well as representations that use knowledge bases like WordNet (Faruqui and Dyer, 2015).

However, these resources are (by definition, as decontextualized judgments) not directly amenable to evaluating how well a model incorporates *context* into its semantic representation of a given word.

**Word Sense Disambiguation**

In Word Sense Disambiguation (WSD), a classifier predicts the "sense" of an ambiguous word in a given context, often using a contextualized embedding. WSD relies on annotated sense labels, which in turn requires determining whether any given pair of word uses belong to the same or distinct senses—i.e., whether to "lump" or "split". There is considerable debate about how *granular* word sense inventories should be (Hanks, 2000; Brown, 2008a);[4] resources range in granularity from WordNet (Fellbaum, 1998) to the Coarse Sense Inventory, or CSI (Lacerra et al., 2020). Recent work using coarse-grained sense inventories has achieved success rates of 85% and beyond (Lacerra et al., 2020; Loureiro et al., 2020).

In terms of the criteria listed above, WSD satisfies the Disambiguation Criterion, but not the Contextual Gradation Criterion. WSD only captures a model's ability to distinguish between distinct senses; it does not assess how meaning is modulated within a given sense category, e.g., that a human comprehender might consider the meaning of *runs* in "the cheetah runs" as more similar to "the jaguar runs" than to "the toddler runs", or that some uses of a sense might be more prototypical than others.

**Contextualized Word Similarity and Relatedness**

There have been several recent efforts to address this gap in the literature:

The Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012) contains similarity judgments for 2,003 English word pairs in a sentence context. Approximately 12% of

[4] This also raises deeper philosophical issues about exactly what qualifies as a "sense" (Hanks, 2000; Tuggy, 1993; Geeraerts, 1993; Kilgarriff, 2007); answering these questions is beyond the scope of this paper.

the pairs contain the same word (e.g., “pack his bags” vs. “pack of zombies”), though not always in the same part of speech; in most cases, the words compared are different (e.g., “left” vs. “abandon”). This dataset is a useful step towards contextualized similarity judgments, but because most pairs contain different words (or the same word in different parts of speech), static word embeddings such as Word2Vec can still perform quite well without considering the context at all (Pilehvar and CamachoCollados, 2018).

The Word in Context (WiC) dataset (Pilehvar and Camacho-Collados, 2018) contains a set of over 7,000 sentence pairs with an overlapping English word, labeled according to whether the use of that word corresponds to same or different senses. As Pilehvar and Camacho-Collados (2018) note, the structure of the dataset requires some form of contextualized meaning representation to perform above a random baseline, which makes it more suitable for interrogating contextualized embeddings. However, the task is a binary classification task along the lines of WSD, making it harder to assess the Contextual Gradation Criterion.

The CoSimLex dataset (Armendariz et al., 2020), created with the Graded Word Similarity in Context (GWSC) task, contains graded similarity judgments for a number of word pairs across English (340), Croatian (112), Slovene (111), and Finnish (24). Each pair of words is rated in two separate contexts, yielding 1174 scores in total. Sentence contexts were extracted from each language’s Wikipedia. Unlike WiC, the word pairs do not actually contain the same word—rather, for any given word pair (e.g., “beach” and “seashore”), there are at least two pairs of sentence contexts with associated similarity judgments. Thus, this dataset can be used to assess graded differences in contextualized meaning representations, but not directly for the *same* ambiguous word.

**Contextualized Similarity of Ambiguous Words**

Finally, one dataset (Haber and Poesio, 2020a,b) contains graded similarity judgments (as well as copredication acceptability judgments) for a number of polysemous words in distinct sentential contexts, meeting both Contextual Gradation and the Disambiguation criteria.

The main limitations of this dataset are its size (it contains examples for only 10 polysemes), as well as the fact that the sentence frames are also not always controlled for each polysemous word.

**Summary**

Most datasets reviewed above allow practitioners to evaluate models on their ability to disambiguate (i.e., the Disambiguation Criterion) or their ability to capture graded differences in word relatedness (i.e., the Contextual Gradation Criterion); one dataset (Haber and Poesio, 2020) meets both criteria. But to our knowledge, no datasets contain graded relatedness judgments for ambiguous words in tightly controlled sentence contexts, with inflection and part-of-speech controlled across each use. In Section 3 below, we describe the procedure we followed for constructing such a dataset.

**RAW-C: Relatedness of Ambiguous Words, in Context**

Items were adapted from stimuli used in past psycholinguistic studies, which contrasted behavioral responses to homonymous and polysemous words, either in isolated lexical decision tasks (Klepousniotou and Baum, 2007) or in a disambiguating context (Klepousniotou, 2002; Klepousniotou et al., 2008; Brown, 2008b). We selected 115 words in total. For each ambiguous word (e.g., "bat"), we created four sentences: two each for two distinct meanings of the word. We attempted to match the sentence frames as closely as possible, in most cases altering only a single word[5] across the four sentences to disambiguate the intended meaning:

[5] There were 13 words for which at least one of the four sentences used a different article ("a" vs. "an"), in addition to having a different disambiguating word.

1. He saw a *fruit* bat.
2. He saw a *furry* bat.
3. He saw a *wooden* bat.
4. He saw a *baseball* bat.

We also labeled each word according to whether the two distinct meanings were judged by lexicographers to be Polysemous or Homonymous. Distinguishing homonymy from polysemy is notoriously challenging (Valera, 2020); common tests include determining whether the two meanings share an etymology (polysemy) or not (homonymy), or determining whether the two meanings are conceptually related (polysemy) or not (homonymy). Both tests can be criticized on multiple grounds (Tuggy, 1993; Valera, 2020), and do not always point in the same direction (e.g., etymologically related words sometimes drift apart, resulting in apparent homonymy). For our annotation, we consulted both the online Merriam-Webster Dictionary (https://www.merriam-webster.com/) and the Oxford English Dictionary, or OED (https://www.oed.com/), and identified whether each dictionary listed the two meanings in question in separate lexical entries (homonymy), or as different senses under the same lexical entry (polysemy).[6] For example, both dictionaries list the *animal* and *meat* senses of the word "lamb" as different senses under the same lexical entry, whereas they list the *animal* and *artifact* senses of the word "bat" under different lexical entries. There was one word ("drill") on which the two dictionaries did not agree; in this case, we labeled the two meanings ("electric drill" vs. "grueling drill") as homonymy (as per the OED).

[6] Our primary goal with this labelling was not to definitively distinguish homonymy from polysemy; as noted above, there is no single, universal criterion for doing so, and different criteria might be more or less relevant for different purposes. It was simply to specify how lexicographers treat the different words, in case that information is useful for users of the resource.

There were also three words for which neither dictionary distinguished the two meanings (either in terms of homonymy or polysemy). For example, "best-selling novel" and "thick novel" refer to *cultural* and *physical* artifacts, respectively, but are not listed as distinct senses. Again, this highlights the challenge of distinguishing outright *ambiguity* from *context-dependence*; these items were included in the annotation study described below, but were excluded from the final set of norms, thus resulting in 112 target words altogether.[7] Each word was used in four sentences, for a total of six *sentence pairs.* 84 of the target words were nouns, and 28 were verbs (note that Part-of-Speech was always held constant *within* each word).

## Human Annotation

### Participants

81 participants were recruited through UC San Diego's undergraduate subject pool for Psychology, Cognitive Science, and Linguistics students. Participants received class credit for participation. Three participants were removed for failing the bot checks at the beginning of the study, and one was removed for failing the catch trials embedded in the experiment, leaving 77 participants in total (59 Female, 16 Male, 2 Non-binary). The median age of participants was 20 (M = 20.22, SD = 2.7), with ages ranging from 18 to 38. 74 participants self-reported as being native speakers of English.

### Materials

[7] The existence of these "Unsure" items, as well as items for which the two dictionaries disagreed on the issue of homonymy vs. polysemy, raises the question of whether empirical measurements such as relatedness judgments (or even cosine distance) could help inform lexicographic decisions. As a proof of concept, we trained a logistic regression classifier (using leave-one-out cross-validation) to predict whether two contexts of use belonged to the Same Sense, using Mean Relatedness. The classifier successfully categorized 86.76% of held-out test items as belonging to the same or different senses. Further, for different sense items only, a trained classifier successfully categorized 79% of held-out test items as polysemous or homonymous. While only a proof of concept, this demonstration suggests a promising avenue for future research.

We used the original set of 115 words described in Section 3, i.e., including the three items labeled “Unsure”. Each word had four sentences; accounting for order, this resulted in twelve possible sentence pairs (six pairs, with two orders each) for each word, for a total 1380 items.

**Procedure**

After giving consent, participants answered two questions designed to filter out bots (e.g., “Which of the following is not a place to swim?”, with the correct answer being “Chair”). They were then given instructions, which included a description of how the meaning of a word can change in different contexts.

On each page of the study, participants were shown a pair of sentences, with the target word bolded (see Figure 1 for an example). They were asked to indicate how related the uses of that word were across the two sentences, with a labeled Likert scale ranging from “totally unrelated” to “same meaning”.

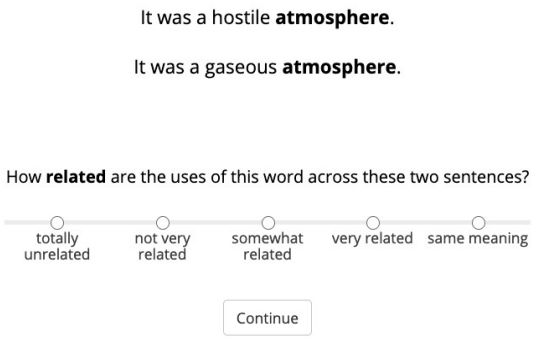


Figure 1: Example item from study.

We included two “catch” trials in the study to identify participants who did not pay attention. In one, the two sentences were identical, such that the correct answer is “same meaning”; the other featured a homonym with two different parts of speech (*rose.v* and *rose.n*), such that the correct answer was “totally unrelated”.

Excluding the catch trials, participants saw 115 sentence pairs total; no word was repeated twice across trials for the same participant. The comparisons any given subject saw for a given word were randomly sampled from the 12 possible sentence pairs, and the order of trials was randomized.[8]

## Analysis and Results

The analyses run below were performed on the 112 target words (i.e., excluding the "Unsure" items). Human annotations were assigned to a scale from 0 ("totally unrelated") to 4 ("same meaning").

### Analysis of Sentence Pairs

Before analyzing the responses of human annotators, we first sought to characterize how well two neural language models captured the *categorical* structure in the dataset—i.e., whether their contextualized representations could be used to distinguish same-sense from different-sense uses of the same word, as well as words labeled as different-sense Homonyms from different-sense Polysemes.

We ran every sentence through two language models: ELMo, using the Python AllenNLP package (Gardner et al., 2017), and BERT, using the bert-embedding package.[9] Then, for each sentence pair, we computed the Cosine Distance between the contextualized representations of the target wordform (e.g., *bat* in "He saw the furry bat" and "He saw the wooden bat"). The distribution of Cosine Distances is visualized in Figure 2.

[8] Based on the suggestion of an anonymous reviewer, we also ran a follow-up norming study to collect estimates of sense frequency bias (sometimes called *dominance*); sense dominance is known to play an important role in the processing of ambiguous words (Klepousniotou and Baum, 2007; Rayner et al., 1994; Binder and Rayner, 1998; Leinenger and Rayner, 2013). These dominance norms are included in the final dataset.

[9] https://pypi.org/project/ bert-embedding/

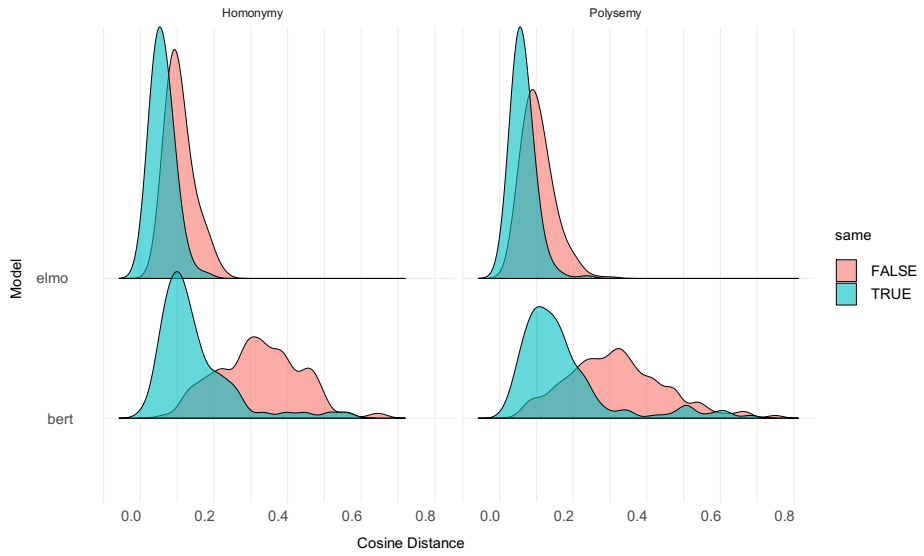


Figure 2: Cosine Distances between the target word's contextualized embeddings for both language models, plotted by Same Sense (True vs. False) and Ambiguity Type (Homonymy vs. Polysemy).

We also performed several statistical analyses, using the lme4 package in R (Bates et al., 2015). In each case, we compared a full model to a reduced model using a log-likelihood ratio test. All models had Cosine Distance as a dependent variable, and included Part-of-Speech as a fixed effect, random intercepts for Word and Language Model (i.e., ELMo vs. BERT), and by-Word random slopes for the effect of Same Sense.

Adding a fixed effect of Same Sense significantly improved model fit [$\chi^2(1) = 143.72, p < .001$], with same-sense uses significantly closer than different-sense uses [$\beta = -.099, SE = 0.005$]. However, adding an interaction between Same Sense and Ambiguity Type (as well as fixed effects of both) did not significantly improve the fit above a model omitting the interaction [$\chi^2(1) = 2.19, p = 0.14$]. In other words, both language models could differentiate same-sense and different-sense uses of an ambiguous word, but their ability to discriminate between Homonymy and Polysemy was marginal at best.

**Analysis of Human Annotations**

Our primary goal was understanding the distribution of human relatedness annotations—both in terms of how it reflects the underlying categorical structure of the dataset (e.g., Homonymy vs. Polysemy), as well as the Cosine Distance measures from each language model. As in the section above, we constructed a series of linear mixed effects models and performed log-likelihood ratio tests for each model comparison; in each case, the dependent variable was Relatedness. All models included a fixed effect of Part-of-Speech, by-subject and by-word random slopes for the effect of Same Sense, by-subject random slopes for the effect of Ambiguity Type, and random intercepts for subjects and items.

First, we asked whether participants' relatedness judgments varied across same-sense and differentsense sentence pairs. We added a fixed effect of Same Sense to the base model described above, along with fixed effects for the Cosine Distance measures from BERT and ELMo. This model explained significantly more variance than a model omitting only Same Sense [$\chi^2(1) = 207.11, p < .001$], with same-sense uses receiving higher relatedness judgments on average [$\beta = 1.94, SE = 0.1$]. The median relatedness judgment for samesense uses was 4 ($M = 3.46, SD = 1.02$), while the median relatedness judgment for differentsense uses was 1 ($M = 1.31, SD = 1.45$).

Second, we asked whether participants' judgments were sensitive to the distinction between Homonymy and Polysemy. We added an interaction between Same Sense and Ambiguity Type (along with a fixed effect of Ambiguity Type) to the model described above. The interaction significantly improved model fit [$\chi^2(1) = 25.45, p < .001$]. The median relatedness for both same-sense homonyms and polysemes was 4, whereas the median relatedness for different-sense homonyms (0) was lower than that for different-sense polysemes

(2). Further, as depicted in Figure 3, there was considerably more variance across polysemous words than homonymous words—this makes sense, given that some polysemous meanings are highly related (e.g., "pet chicken" vs. "roast chicken"), while others are more distant (e.g., "desperate act" vs. "magic act").

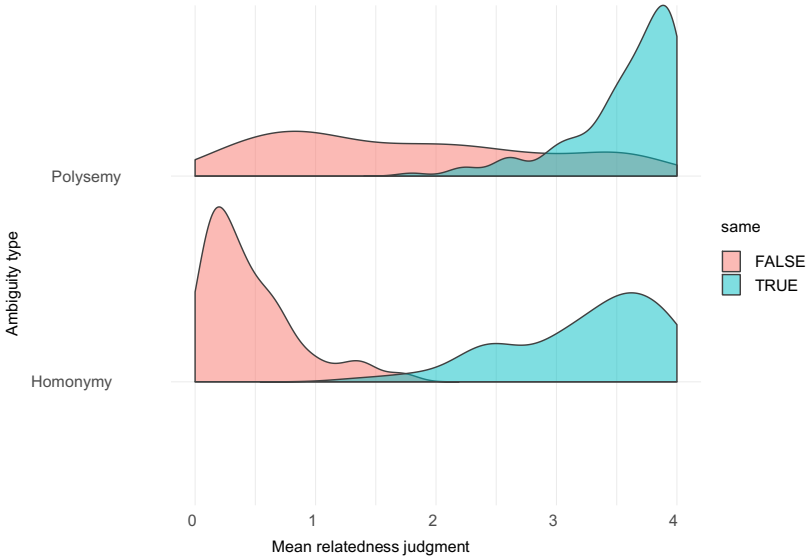


Figure 3: Mean relatedness judgments for each sentence pair, plotted by by Same Sense (True vs. False) and Ambiguity Type (Homonymy vs. Polysemy).

Third, we asked whether the Cosine Distance measures explained independent variance above and beyond that explained by Same Sense and Ambiguity Type. A full model including all factors explained more variance than a model excluding only the Cosine Distance measure from BERT [$\chi^2(1) = 36.19, p < .001$], as well as a model excluding only the Cosine Distance measure from ELMo [$\chi^2(1) = 16.92, p < .001$]. This indicates that Relatedness does not vary purely as a function of the categorical structure in the dataset—the graded relatedness judgments were sensitive to subtle differences in context.

**Inter-Annotator Agreement**

Inter-annotator agreement was assessed by calculating the average Spearman's rank correlation between each participant's responses and the Mean Relatedness for the set of 112 items observed by that participant—where Mean Relatedness was calculated after omitting responses by the participant in question. This answers the question: to what extent do each participant's responses correlate with the consensus rating by the 76 other participants? Using this method, the average correlation was $\rho = 0.79$, with a median of $\rho = 0.81$ ($SD = .07$). The lowest agreement was $\rho = 0.55$, and the highest was $\rho = 0.88$.

**Evaluation of Language Models**

To evaluate the language models, we collapsed across the single-trial data and computed the Mean and Median Relatedness for each unique sentence pair. The distribution of Mean Relatedness judgments is depicted in Figure 3.

As in past work (Hill et al., 2015), we computed the Spearman's rank correlation between the distribution of Cosine Distance measures (from each model) and the Mean Relatedness for a given sentence pair. BERT performed slightly better than ELMo ($BERT_\rho = -0.58, ELMo_\rho = -0.53$).[10] Putting this in context, both models performed considerably worse than the average inter-annotator agreement score ($\rho = 0.79$).

We also computed the $R^2$ of a linear regression including the Cosine Distance measures from *both* BERT and ELMo. Combined, both measures explained roughly 37% of the variance in Mean Relatedness judgments ($R^2 = 0.37$). Surprisingly, this was only slightly more than half the variance explained by a linear regression including only the interaction between Same Sense and Ambiguity Type ($R^2 = 0.66$), as well as a regression including all factors ($R^2 = 0.71$).

[10] Note that larger values of Cosine Distance indicate a *larger* distance between two vectors; thus, a negative correlation is expected between relatedness and Cosine Distance.

By visualizing the residuals from the linear regression with only BERT and ELMo (see Figure 4), we see that Cosine Distance appears to systematically *underestimate* how related participants find same-sense judgments to be (for both Polysemy and Homonymy). Further, we see that Cosine Distance systematically *overestimates* how related participants find different-sense Homonyms to be.

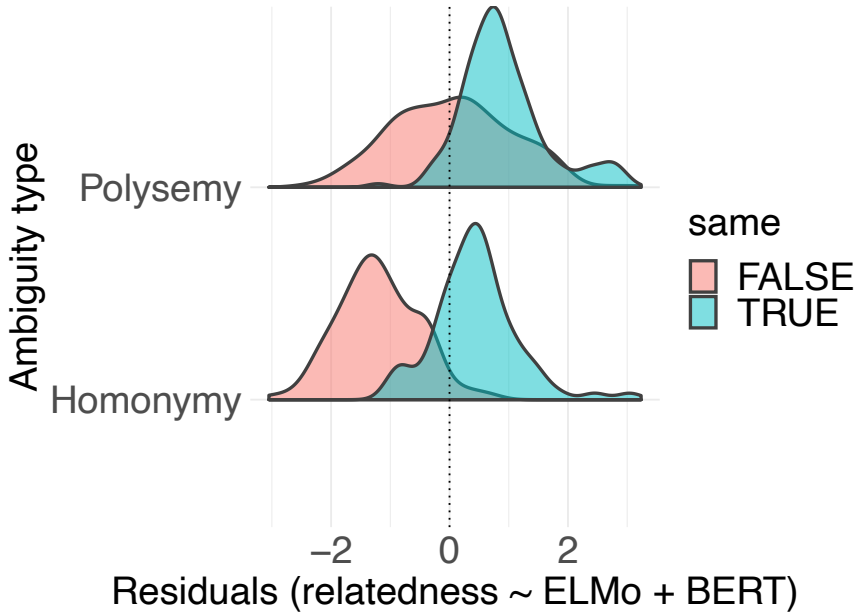


Figure 4: Residuals of a linear regression including Cosine Distance measures from both BERT and ELMo, plotted by by Same Sense (True vs. False) and Ambiguity Type (Homonymy vs. Polysemy).

**Discussion**

Word meanings are dynamic, dependent on the contexts in which those words appear—and some words are even ambiguous, generating distinct, incompatible interpretations in different situations (e.g., "fruit *bat*" vs. "baseball *bat*").

RAW-C contains graded relatedness judgments (by human annotators) for ambiguous English words in distinct sentential contexts. Importantly, the ambiguous wordform (e.g., "bat") is always matched for both part-of-speech and inflection across each sentence pair; 84 of the

target words are nouns, and 28 are verbs. Each word has relatedness judgments for six different sentences pairs (four unique sentences): two same-sense pairs, and four different-sense pairs. Same sense pairs convey the same meaning, according to Merriam-Webster and the OED (e.g., "fruit bat" and "furry bat"), while different sense pairs correspond to meanings listed in either distinct lexical entries (e.g., "fruit bat" and "wooden bat") or distinct sub-entries (e.g., "marinated lamb" and "baby lamb"). Furthermore, different-sense pairs are labeled according to whether they are related via homonymy or polysemy, a relevant distinction for both lexicographers and psycholinguists—recent evidence suggests that polysemous and homonymous meanings are represented differently in the mental lexicon (Klepousniotou, 2002; Klepousniotou and Baum, 2007). Finally, the sentential context is always tightly controlled; in most pairs, only one word differs across the two sentences (e.g., "fruit" vs. "furry").

In Section 5, we reported several primary findings. First, contextualized representations from both BERT and ELMo capture the distinction between same-sense and different-sense uses of a word, but their ability to distinguish between homonymy and polysemy is marginal at best. This contrasts with other recent work (Nair et al., 2020), suggesting that BERT *is* able to differentiate between homonymy and polysemy. One possible explanation for this difference in results is that Nair et al. (2020) used naturally-occurring sentences from Semcor (Miller et al., 1993), whereas our sentence contexts were more tightly controlled. Our results indicate that even the presence of a single disambiguating word can trigger nuanced differences in semantic representation in humans, but not necessarily in current neural language models. Second, we found that both BERT and ELMo explain independent sources of variance in human relatedness judgments, above and beyond Same Sense and Ambiguity Type (i.e., homonymy vs. polysemy). This is encouraging, because it demonstrates a direct benefit of graded (rather than categorical)

judgments; for example, among the broad category of different-sense polysemous pairs, some are closely related (e.g., “marinated lamb” and “baby lamb”), and others are considerably less closely related (e.g., “hostile atmosphere” and “gaseous atmosphere”). Overall, contextualized embeddings from BERT were better at predicting human relatedness judgments than those from ELMo—this is consistent with past work (Wiedemann et al., 2019) suggesting that BERT outperforms ELMo on tasks involving sense disambiguation.

Importantly, however, both BERT and ELMo failed to capture variance in relatedness judgments that is captured by Same Sense and Ambiguity Type. As depicted in Figure 4, Cosine Distance tended to *underestimate* how related humans find same-sense uses to be, and *overestimate* how related humans find different-senses to be. This is not entirely surprising, given that neither BERT nor ELMo are equipped with discrete sense representations—at most, they produce contextualized embeddings that are amenable to supervised classification or unsupervised clustering. Yet this also illustrates that—at least on this task—humans *do* appear to draw on some manner of (likely fuzzy) categorical representation, such that the difference between two contexts of use is *compressed* for same-sense meanings, and *exaggerated* for different-sense meanings (particularly for homonyms). This suggests several exciting avenues for future work: can neural language models such as BERT be augmented with semantic knowledge or representational schemes that improve their performance on RAW-C or similar tasks? Both possibilities are explored in Section 6.1 below.

## Future Work

As Bender and Koller (2020) note, most language models are trained on linguistic form alone. In contrast, human language knowledge is *grounded* in our embodied experience of the world (Bisk et al., 2020). To the extent that human sense representations are guided by distinct sensorimotor

or social-interactional associations, equipping language models with this information ought to facilitate their ability to distinguish between distinct meanings of a word (i.e., the Disambiguation Criterion) and modulate a given meaning in context (i.e., the Contextual Gradation Criterion).

Practitioners could also look to (and in turn, inform) models of the human mental lexicon (Nair et al., 2020). Several promising models attempt to address the *continuous* nature of word meaning, as well as the issue of apparent category boundaries (i.e., word senses) (Rodd et al., 2004; Elman, 2009); at this stage, the role of continuity vs. categorical structure in human sense representations remains an open question. Models such as SenseBERT (Levine et al., 2020) incorporate high-level sense knowledge into internal representations from the beginning, and find improvements on several WSD tasks—would this approach, or others like it, yield an improvement on RAW-C as well?

## Limitations of Dataset

RAW-C has multiple limitations, some of which could also be addressed in future work. First, the broad category of "polysemy" is often subdivided into different mechanisms or manners of conceptual relation, such as metaphor and metonymy. This distinction is also believed to be cognitively relevant, with some evidence that metaphorically related senses are represented differently than metonymically related ones (Klepousniotou, 2002; Klepousniotou and Baum, 2007; Lopukhina et al., 2018; Yurchenko et al., 2020). Future work could annotate polysemous word pairs for whether they are related by metaphor, metonymy, or another class of semantic relation—annotations could even be made as granular as the specific semantic relation involved (e.g., Animal for Meat) (Srinivasan and Rabagliati, 2015). This finer-grained coding could be used to assess exactly which kinds of semantic relation correlate with the distributional profile of

word tokens—i.e., are accessible from linguistic form alone—and which require some external module, whether in the form of grounded world knowledge or a structured knowledge base.

Another possible limitation is the fact that RAWC contains experimentally controlled minimal pairs, instead of naturally-occurring sentences (Nair et al., 2020; Haber and Poesio, 2020a,b). On the one hand, naturalistic sentences are useful for evaluating models on WSD "in the wild" (and indeed, there are a number of useful datasets for this purpose). On the other hand, controlled datasets are useful if one's goal is to better understand a particular model or linguistic phenomenon— especially if this also allows a direct comparison with human annotations. For example, our analyses suggest that human sense representations must involve some additional levels of processing or information beyond the statistical regularities in word co-occurrence captured by BERT and ELMo. Moving forward, we hope that experimentally controlled datasets such as RAW-C will serve as a useful complement to existing, more naturalistic datasets.

## Conclusion

We have presented a novel dataset for evaluating contextualized language models: RAW-C (Relatedness of Ambiguous Words, in Context). This resource contains both categorical annotations, derived from expert lexicographers (MerriamWebster and the OED), as well as graded relatedness judgments from human participants. We found that contextualized representations from BERT and ELMo captured some variance ($R^2 = .37$) in these relatedness judgments, but that the distinction between same-sense and different-sense uses, as well as between homonymy and polysemy, explains considerably more ($R^2 = .66$). Finally, we argued that this gap in performance represents an exciting opportunity for further development, and for crosspollination between experimental psycholinguistics and NLP.

**Acknowledgments**

Chapter 2, in full, is a reprint of the material as it appears in the Proceedings of the 59th Annual Meeting of the Association for computational Linguistics and the 11th International Joint Conference on Natural Language Processing, in 2021 (August). Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

# CHAPTER 3: CONTEXTUALIZED SENSORIMOTOR NORMS

Most large language models (LMs) are trained on linguistic input alone. This approach may be fundamentally limited when it comes to language understanding (Bender and Koller, 2020; Bisk et al., 2020; Tamari et al., 2020), as the meaning of a word arguably depends on factors beyond which words it co-occurs with. In particular, humans appear to *ground* a word's meaning in a rich network of sensorimotor associations (Pulvermüller, 1999; Bergen, 2012; Bergen and Feldman, 2008; Barsalou, 1999; Winter and Bergen, 2012; Barsalou, 2008; Glenberg and Kaschak, 2002). For example, our understanding of the word "table" incorporates not just the words that frequently co-occur with "table", but also our embodied experience of tables: how they look, how they feel, which parts of our body we use to interact with them, and more. If human-like language understanding depends on grounding words in non-linguistic associations (Harnad, 1990), then LMs trained on text alone will never reach human levels of understanding (Bender and Koller, 2020).

One promising solution is to use human judgments of a word's sensorimotor associations, such as the Lancaster Sensorimotor Norms (Lynott et al., 2019) (hereafter LS Norms), to help *ground* LM representations, as well as *evaluate* the extent to which those representations capture sensorimotor properties of word meaning. The LS Norms provide human ratings about the extent to which an isolated word (e.g., "table") is strongly associated with various *sensory modalities* (e.g., Vision vs. Touch) and *action effectors* (e.g., Hand/Arm vs. Foot/Leg). Recent work (Kennington, 2021; Wan et al., 2020b,a) has found that integrating these norms improves the performance of language models on several NLP tasks, such as GLUE (Wang et al., 2018) and metaphor detection (Wan et al., 2020a).

Despite the promise and early success of this approach, it faces a key limitation: resources like the LS Norms typically contain just a single set of judgments for each word. In practice, however, many words are *ambiguous* (Rodd et al., 2004; Haber and Poesio, 2021). In English, anywhere from 7% (Rodd et al., 2004) to 15% (Trott and Bergen, 2020) of words have multiple, unrelated meanings—and as many as 84% are polysemous, i.e., they have multiple, related meanings (Rodd et al., 2004). For example, the word "table" may refer to a piece of furniture or to a database organized into rows and columns. Further, even very similar uses of a word, like "lemon", in its fruit-denoting sense, evoke different sensorimotor associations in different contexts (e.g., "She peeled the lemon" vs. "She put the lemon in the bag") (Yee and Thompson-Schill, 2016; Elman, 2009; Trott et al., 2020). Accordingly, there is evidence that ratings of sensorimotor strength or concreteness can vary considerably depending on whether a word is presented alone or in context (Scott et al., 2019), or as a function of which context a word is presented in (Reijnierse et al., 2019). This suggests that any effort to *ground* words should account for the fact that most words are ambiguous, with dynamic, context-sensitive meanings subject to construal. Further, attempts to *evaluate* grounded language models must consider not only how well they capture the sensorimotor properties of a word in isolation, but also how successfully they capture context-dependent variation in a word's sensorimotor profile.

In Section 2, we first describe related resources, as well as work on grounding large LMs using psycholinguistic resources and multimodal input. In Section 3, we introduce the Contextualized Sensorimotor Norms (CS Norms), a dataset of sensorimotor judgments about ambiguous words in context. In Section 4, we provide descriptive statistics about the CS Norms, as well as comparisons to other factors such as the *dominance* of a particular sense. In Section 5, we show that a metric derived from the CS Norms—the Sensorimotor Distance between two

contexts of use—improves our ability to predict contextualized relatedness judgments, above and beyond a similar metric derived from BERT (Devlin et al., 2019). Finally, in Section 6, we discuss limitations of these norms, as well as avenues for future research.

## Related Resources

There are a number of existing lexical resources with information about the concreteness or sensorimotor strength of words (Coltheart, 1981; Brysbaert et al., 2014b). For example, the Brysbaert concreteness norms contain concreteness judgments for approximately 37,000 English words (Brysbaert et al., 2014b); concreteness ratings have also been collected for Dutch (Brysbaert et al., 2014a), Croatian (Coso et al.′ , 2019), and more.

Judgments of concreteness or overall sensorimotor strength are limited in that they do not account for which sensorimotor features are particularly salient. More recently, researchers have collected ratings about multiple semantic features for each word, including its sensorimotor associations (Lynott et al., 2019), as well as even more fine-grained judgments within each modality (e.g., for Vision, whether the referent is Fast or Slow; for Touch, whether it is Hot or Cold) (Binder et al., 2016). Of these, the largest dataset is the Lancaster Sensorimotor Norms (Lynott et al., 2019), which includes 11-dimensional judgments for about 40,000 English words. This approach has been extended to other languages, such as French (Miceli et al., 2021) and Dutch (Speed and Brybaert, 2021). Again, in each case, the words were presented without context.

Finally, several datasets have collected concreteness judgments about words in context (Scott et al., 2019; Reijnierse et al., 2019). However, to our knowledge, no dataset includes judgments about *which* sensorimotor features are particularly salient in different linguistic contexts.

### Grounding LMs with Psycholinguistic Resources

Recent work in NLP has begun to incorporate these psycholinguistic resources. One approach attempts to predict these judgments about concreteness or salient sensorimotor features from LM representations, with varying degrees of success (Thompson and Lupyan, 2018; Turton et al., 2020; Chersoni et al., 2020; Utsumi, 2020). Another approach uses sensorimotor features to augment the ability of an LM on an applied task, such as the GLUE benchmark (Kennington, 2021) or metaphor detection (Wan et al., 2020b). As mentioned in Section 1, these experiments are limited in that the sensorimotor features themselves were obtained for words in isolation.

### Grounding LMs with Multimodal Input

An alternative approach is to ground LM representations more directly in multimodal input. Most of this work has emphasized the visual modality, linking words to static images (Kiros et al., 2018; Su et al., 2020) or video (Zellers et al., 2021b). This paradigm shows considerable promise, though it is naturally limited by resource constraints; obtaining reliable multimodal data and aligning it to language can be both time-consuming and costly.

### Summary

There is considerable interest in *grounding* among both psycholinguists and NLP practitioners. To that end, psycholinguists have developed large linguistic resources, which some NLP researchers have used to improve LMs.

Still, one limitation of the majority of existing resources is that they do not contain judmgents about different sensorimotor features for words in different contexts. Because most words are ambiguous, this makes it difficult to know which meaning the sensorimotor judgments reflect, which in turn reduces the precision and utility of these resources.

## Contextualized Sensorimotor Norms

Our primary goal was to collect sensorimotor judgments about ambiguous words, appearing in controlled sentential contexts. We used sentences from the RAW-C (Relatedness of Ambiguous Words–in Context) dataset (Trott and Bergen, 2021). RAWC contains relatedness judgments for 672 English sentence pairs, each containing the same target word (e.g., "bat") in either the same meaning (e.g.,
"furry bat" vs. "fruit bat") or different meaning (e.g., "furry bat" vs. "wooden bat"); it also contains dominance judgments about the relative salience of each meaning. There were 448 unique sentences in total (112 target words, with 4 sentences each).

We collected judgments about the sensorimotor associations evoked by the target word in each of the 448 sentences. This provided a more direct analogue to the Lancaster Sensorimotor Norms (Lynott et al., 2019), in which participants observed a particular lexical item (e.g., "bat") and provided ratings about its associated sensory modalities (e.g., Vision) or action effectors (e.g., Hand/Arm).

### Participants

Our goal was to collect a minimum of 10 judgments per sentence. Thus, we recruited participants until each sentence had at least 10 observations, after applying the exclusion criteria.

A total of 377 participants were recruited through UC San Diego's undergraduate subject pool for Psychology, Cognitive Science, and Linguistics students. Participants received class credit for participation. After excluding non-native speakers of English, participants who failed to pass the bot checks, and participants whose inter-annotator agreement score was sufficiently low, we were left with 283 participants. Of these, 223 identified as female (47 male, 8

nonbinary, and 5 preferred not to answer). The mean self-reported age was 20.4 (median = 20, SD =

2.98), and ranged from 18 to 43.

**Procedure**

We adapted the procedure directly from Lynott et al. (2019), with the main modification being that participants now saw words in sentential contexts. Participants were randomly assigned to one of two Judgment Types: 1) Perception, in which they provided ratings about a word's associated sensory modalities (Vision, Hearing, Touch, Interoception, Smell, and Taste); and 2) Action, in which they rated a word's associated action effectors (Hand/Arm, Foot/Leg, Mouth/Throat, Head, and Torso). In total, 132 participants were assigned to the Perception Judgment Type, and 151 were assigned to the Action Judgment Type.

After giving consent, participants answered two bot check questions. They were then told that they would read a series of sentences, each containing a bolded word (e.g., "It was a wooden table"), and that their task was to rate the degree to which they experienced the concept denoted by that word with either six sensory modalities (in the Perception Judgment Type) or five action effectors (in the Action Judgment Type). Ratings ranged from 0 (not at all experienced with that sense/effector) to 5 (experienced greatly with that sense/effector).

Each participant rated approximately 60 sentences overall, randomly sampled from the set of 448 sentences. No participant saw the same target word twice. On each trial, the sentence was displayed at the top of the page, with the target word bolded. Underneath the sentence, the instructions read: "To what extent do you experience WORD:" (for Perception) or "To what extent do you experience WORD by performing an action with the:" (for Action), where "WORD" was replaced with the target word. Underneath the instructions were six (for

Perception) or five (for Action) rating scales, corresponding to each possible sensory modality or action effector. For the Action Judgment Type, the scale was accompanied by a labeled diagram of the body, as in Lynott et al. (2019).

To reach the target of 10 respondents to each word in both Action and Perception tasks, we collected data in two stages. In the first stage (Group 1), participants were randomly assigned to either the Perception or Action Judgment Types, and the sentences they observed were randomly sampled from the set of possible sentences for each word. After we had collected responses from 264 participants in this way, there were still a number of sentences that had very few observations, simply by chance—as well as many with more than ten observations. Thus, in the second stage (Group 2), participants were assigned a mix of Low-N (sentences with fewer than 10 ratings) and High-N (sentences with 10 or more ratings) items. The goal was to speed data collection; to control for potential differences across groups, we compared their distributions of inter-annotator agreement scores, and found no evidence that the different data collection procedures induced different response behavior.

Finally, after providing ratings, participants reported their self-identified gender and age, as well as whether or not they were a native speaker of English.

The data collection was conducted online using JsPsych (De Leeuw, 2015).

**Inter-Annotator Agreement**

We sought to establish the degree to which different participants agreed about their ratings for each sentence, both to characterize the dataset and to exclude participants whose ratings diverged substantively from the rest of the sample. Following past work (Trott and Bergen, 2021), we used a leave-one-out scheme: for each participant, we computed the

Spearman's rank correlation between that participant's responses and the mean ratings for those items from the rest of the sample (excluding the participant's ratings).

Importantly, we did this in two stages. First, we computed the distribution of agreement scores for the 264 participants in Group 1, i.e., the participants for whom each sentence was truly randomly sampled from the set of 448 sentences. Based on this distribution of inter-annotator agreement scores, we excluded a total of 18 participants, whose scores were more than two standard deviations below the mean for that Judgment Type. Among the final set of 246 participants in this group, the mean inter-annotator rank correlation was 0.47 for Action judgments (SD = 0.1) and 0.64 for Perception judgments (SD = 0.11).

Then we considered the 39 participants from Group 2, who provided ratings for a restricted set of sentences, i.e., sentences which either had below 10 judgments from Group 1 (low-N) or had more than 10 judgments from Group 1 (high-N). For each participant in Group 2, we compared the ratings for the high-N items to the mean response for those items among Group 1. After excluding participants with low inter-annotator agreement, we were left with a total of 37 participants in Group 2. The mean rank correlation was 0.5 for Action Judgments (SD = 0.11) and 0.64 for Perception judgments (SD = 0.1).

Finally, we combined the set of inter-annotator agreement scores from both groups, and constructed a linear regression with Rank Correlation as the dependent variable, and main effects of Judgment Type (Action vs. Perception) and Group (Group 1 vs. Group 2), as well as their interaction. There was no significant difference in agreement across groups ($p > .1$), but agreement was significantly higher for Perception ratings than Action ratings [$\beta = 0.17, SE = 0.01, p < .001$].

### Creating the Norms

Once we had obtained a minimum of ten ratings per sentence (per judgment type), we averaged across these ratings to produce a mean and standard deviation for each dimension. For example, the sentence "He saw the furry bat" would contain the mean (and standard deviation) of judgments about the salience of each sensorimotor feature.[11]

## Characterizing the Contextualized Sensorimotor Norms

Our first goal was to characterize the Contextualized Sensorimotor Norms (CS Norms). The norms provide an 11-dimensional vector for each sentential context in which a word appears: the mean sensorimotor strength for 11 dimensions (6 sensory modalities, and 5 action effectors) for a target word in a given context.

### Comparing Sensorimotor Dimensions

As a first step, we visualized the distribution of sensorimotor judgments for each dimension (see Figure 5). Consistent with the original LS Norms (Lynott et al., 2019) and work on the English lexicon more generally (Majid, 2020), judgments tended to be highest for the Vision dimension, and lowest for Olfaction and Taste.

[11] The norms (along with the individualized responses, analysis code, and a Data Sheet) can be found on GitHub: https://github.com/seantrott/cs_norms.

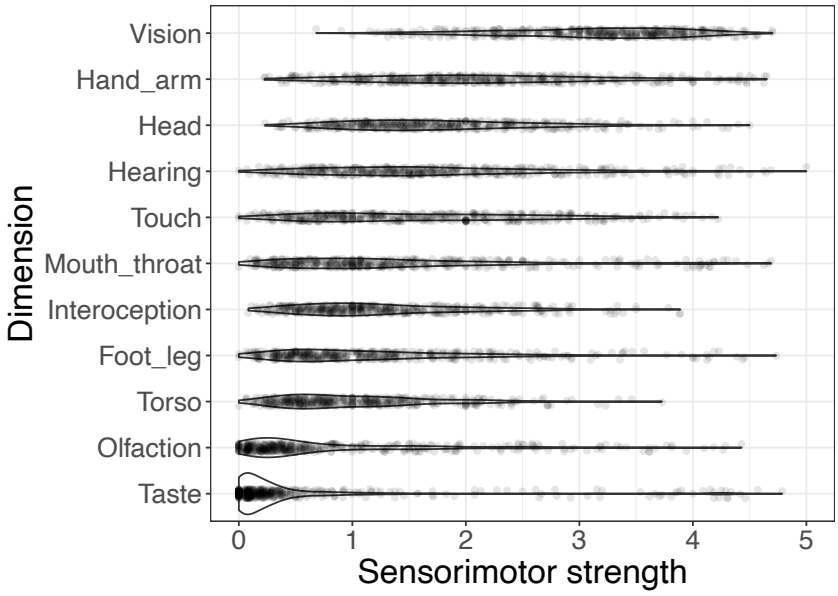


Figure 5: Distribution of mean sensorimotor strength judgments for each dimension.

We then asked which dimensions were correlated with which other dimensions. Consistent with past work (Lynott et al., 2019), we found particularly strong positive correlations between Olfaction and Taste, as well as Foot/Leg and Torso (see Figure 6).

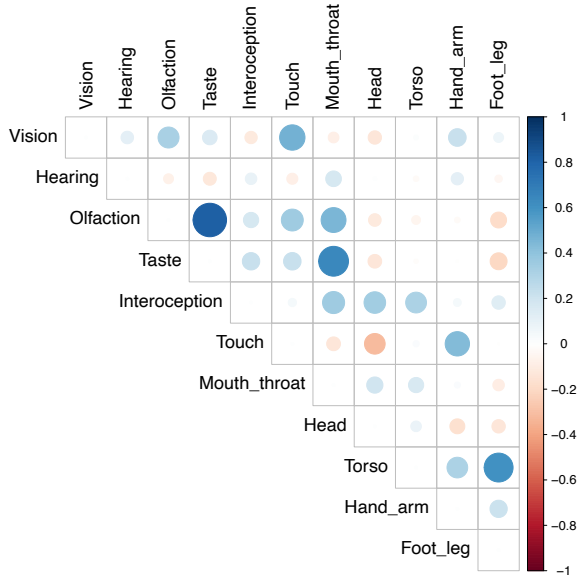


Figure 6: Pearson's correlation coefficients between the strength of each dimension.

**Variance Across Contexts**

A key motivation for the CS Norms was to account for potential variation within each word in terms of which sensorimotor features were most salient across distinct sentential contexts.

We first quantified this variation by normalizing the sensorimotor features for each context of use to the mean norms for that word from the LS Norms. For example, the LS norms have a single 11-dimensional vector for the word “market”; for each of the four sentential contexts in which “market” appeared, we calculated the difference in mean ratings across our norms and the LS Norms. This provides an estimate of the degree to which the human judgments were impacted by the sentential context, as opposed to a representation of the word’s meaning in isolation (as in the LS Norms).

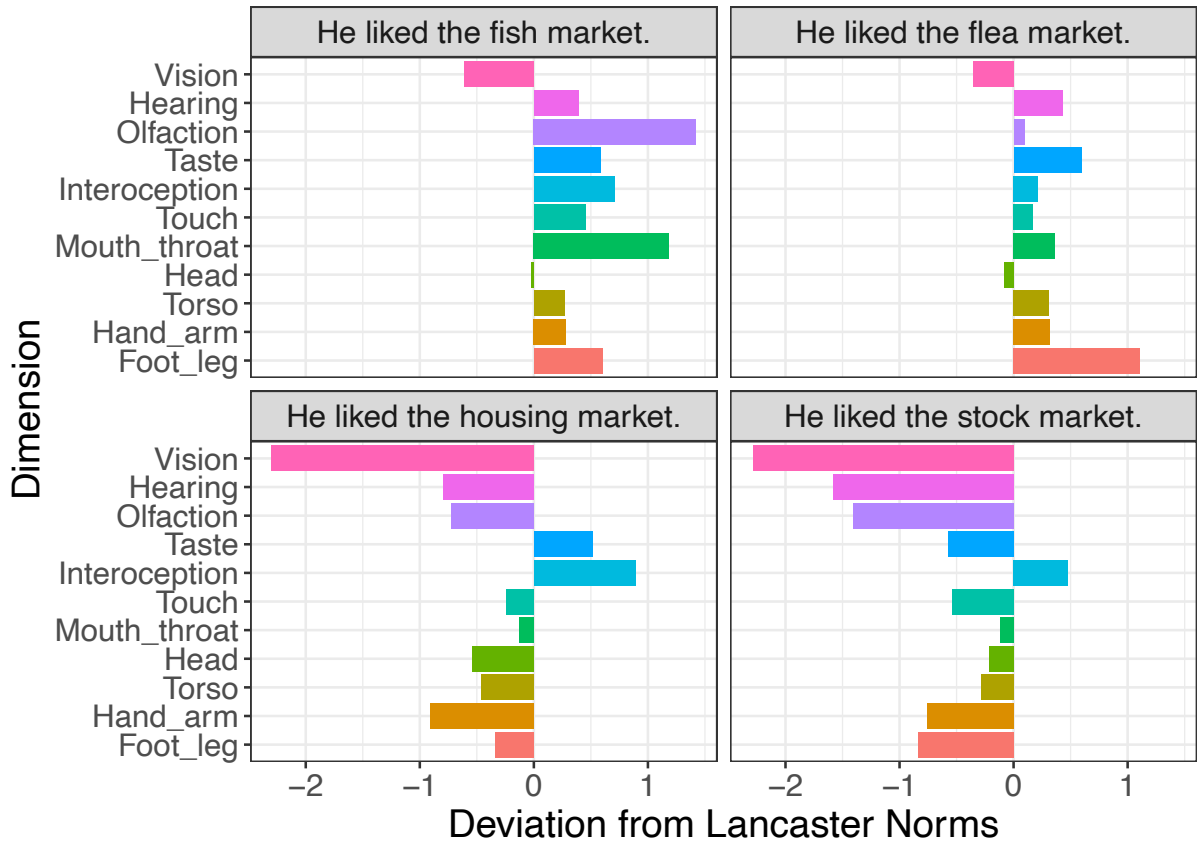


Figure 7: Deviation from the Lancaster Sensorimotor Norms for a specific word, "market", faceted by the distinct sentential contexts in which the word appears.

Figure 7 depicts these deviations from the LS Norms for a specific word, “market”. This word was chosen because it displayed particularly high variation in its overall sensorimotor strength across contexts. The deviations from the LS Norms appear to track the two senses of the word being profiled. The two sentences corresponding to the *location* sense of “market” (i.e., “fish market” and “flea market”) appeared to be closer to the LS Norms (i.e., the deviations were smaller on average); the notable exceptions were the *Olfactory* and *Mouth/Throat* dimensions for the “fish market” context, and the *Foot/Leg* dimension for the “flea market” context. In contrast, the sentences corresponding to the *financial* sense of “market” (i.e., “housing market” and “stock market”) were considerably lower in sensorimotor strength across almost all dimensions,

especially *Vision*. This makes sense, given that this meaning is more metaphorical or abstract than the *location* meaning of "market": apart from representations of their performance, neither housing markets nor stock markets can be visually perceived in the way that fish markets and flea markets can.

**Sense Dominance and Deviation from the Lancaster Norms**

One well-documented property of ambiguous words is that their multiple meanings are not always balanced: one sense is sometimes more cognitively salient than the other. This is called *sense dominance*. The degree of dominance is known to play an important role in the processing of ambiguous words, particularly for homonyms: empirical evidence suggests that comprehenders almost always activate the more dominant sense of a homonym, even when the linguistic context supports the subordinate meaning (Rayner et al., 1994; Binder and Rayner, 1998; Duffy et al., 1988). We used the measure of Sense Dominance in the RAW-C dataset to answer two additional questions about how and why the CS Norms deviate from the decontextualized LS Norms.

First, do the decontextualized LS Norms primarily reflect the more dominant meaning of an ambiguous word? To answer this, we calculated the cosine distance between the decontextualized LS Norm for each word and the sensorimotor norms for that same word in context. We called this Distance to Lancaster. Using the *lme4* package (Bates et al., 2015) in R, we fit a linear mixed effects model predicting Distance to Lancaster, with Dominance as a fixed effect (and random intercepts for words); this model explained more variance than a model omitting only Dominance [$\chi^2(1) = 10.93, p = .001$]. More dominant senses were closer to the

decontextualized norm on average [$\beta = -0.01, SE = 0.003, p = .001$], consistent with the prediction that the LS Norms are more influenced by properties of the dominant sense.

Second, do more dominant senses have stronger or weaker sensorimotor ratings, on average, than the decontextualized rating for that word? For each sentential context, we computed the average difference between each dimension and the corresponding LS Norms, such that a positive value reflects a *more* concrete context. A model predicting this measure was significantly improved by the addition of Dominance [$\chi^2(1) = 38.24, p < .001$]. Dominant senses tended to have stronger sensorimotor ratings, on average, than the decontextualized ratings for that same word [$\beta = 0.09, SE = 0.01, p = .001$].

**Sense Dominance and Sensorimotor Strength**

We also sought to replicate previous work suggesting that more dominant meanings tend to be more concrete (Gilhooly and Logie, 1980). Following Lynott et al. (2019), we created a composite variable called Contextualized Sensorimotor Strength, which measured the maximum strength across the 11 sensorimotor features for each context of use.

Then, we built a linear mixed effects model with Dominance as a dependent variable, fixed effects of Contextualized Sensorimotor Strength, random intercepts for each word, and two covariates reflecting the decontextualized sensorimotor strength for each *word* (i.e., from the Lancaster Sensorimotor Norms dataset). The full model explained significantly more variance than the same model omitting only Contextualized Sensorimotor Strength [$\chi^2(1) = 18.38, p < .001$]. Consistent with past work (Gilhooly and Logie, 1980), contexts of use with higher sensorimotor strength were also rated as more dominant [$\beta = 0.26, SE = 0.06, p < .001$]. (Of course, this finding does not explain *why* more concrete meanings are more dominant than

meanings with less sensorimotor strength; it could be driven by correlations with meaning frequency or even age of acquisition (Gilhooly and Logie, 1980).)

**Sensorimotor Distance**

Another question concerns the relationship *between* contexts of use. Each context of use for a given wordform is associated with its own sensorimotor norms, i.e., the mean ratings for each sensorimotor dimension for a given context; because of this, the similarity or dissimilarity between these contexts can be quantified by calculating the cosine distance between these vectors (Wingfield and Connell, 2021). Thus, we calculated the cosine distance—referred to here as the Sensorimotor Distance—between the vectors corresponding to each sentence pair for each word (672 sentence pairs total). Larger distances reflect more dissimilar contexts of use, while smaller distances reflect more similar contexts. Note that this includes comparisons between contexts corresponding to the same sense and those corresponding to different senses.

We then asked whether Sensorimotor Distance was correlated with other psychologically relevant features, such as whether the two contexts of use corresponded to the same sense or different senses (i.e., Sense Boundary). Based on the preliminary findings in Section 4.2, we predicted that different sense uses would have less similar sensorimotor features.

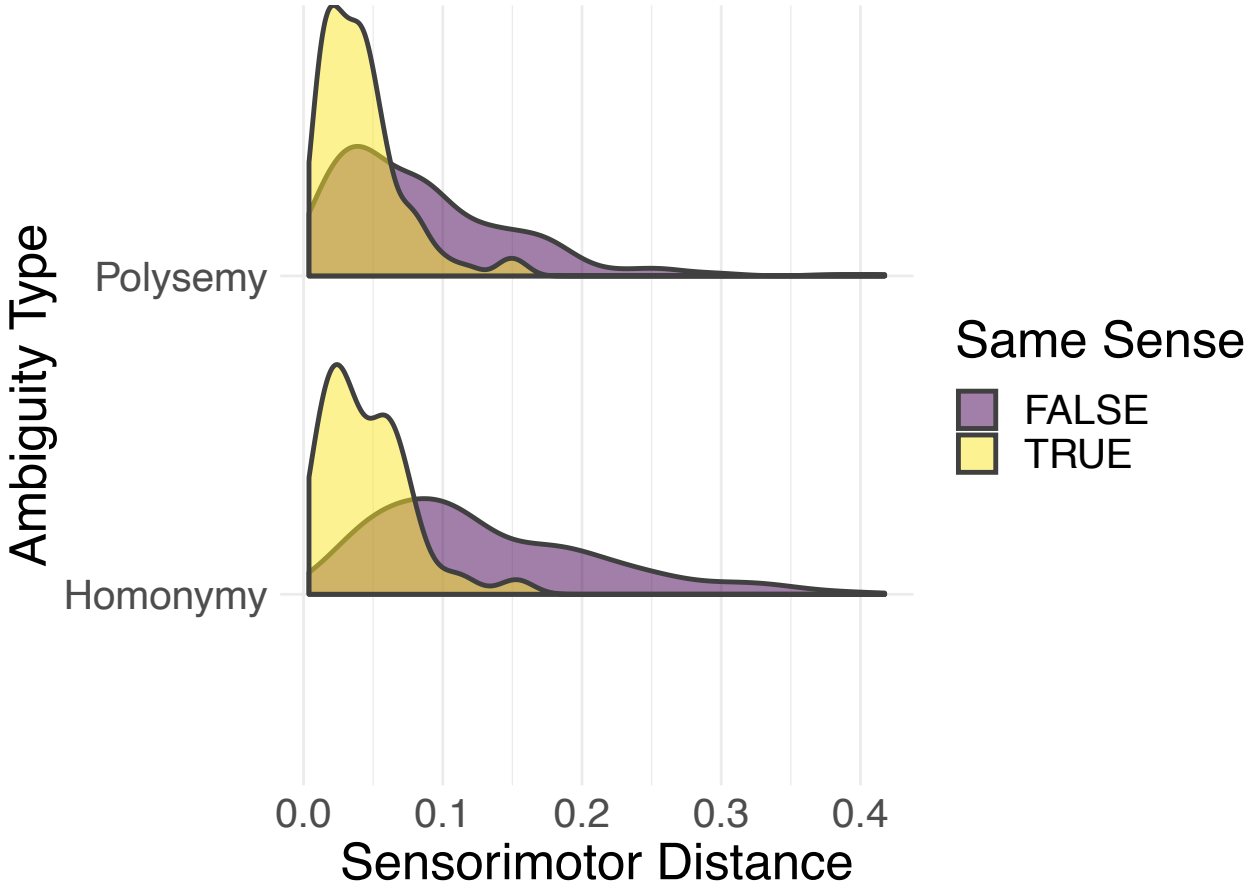


Figure 8: Distribution of sensorimotor distances as a function of same/different sense, as well as the type of ambiguity. Same sense uses have more similar sensorimotor profiles than different sense contexts.

Indeed, as depicted in Figure 8, Sensorimotor Distance was considerably larger for Different Sense than Same Sense contexts. The addition of Sense Boundary to a mixed effects model predicting Sensorimotor Distance improved model fit beyond a model with only Distributional Distance and Ambiguity Type (and random intercepts for words) [$\chi^2(1) = 34.86, p < .001$]. This is also consistent with Figure 3, in which the two *location* senses of "market" were more similar to each other than either was to the two *financial* senses.

## Predictive Utility

We were also interested in the predictive utility of the information provided by the CS Norms, above and beyond other commonly used factors. To what extent do these contextualized ratings encode information that large language models (e.g., BERT) or decontextualized sensorimotor norms (e.g., LS Norms) fail to capture?

We sought to predict the *relatedness* of sentence pairs. RAW-C contains judgments of sense relatedness for each unique sentence pair within each of the 112 words, with a total of 672 sentence pairs (Trott and Bergen, 2021). It is also annotated for whether the two contexts of use correspond to the same or different sense (Sense Boundary), and whether the relationship type is one of homonymy or polysemy (Ambiguity Type). Additionally, past work (Trott and Bergen, 2021) found that relatedness was negatively correlated with the cosine distance between BERT's contextualized embeddings for the target word in each sentence; here, we call this measure the *Distributional Distance*.

We asked whether a linear mixed effects model equipped with those previous factors (Distributional Distance,[12] Sense Boundary, Ambiguity Type, and their interaction, as well as random intercepts for words) could be improved by the addition of Sensorimotor Distance. Indeed, Sensorimotor Distance significantly improved model fit [$\chi^2(1) = 36.74, p < .001$]. As expected, Sensorimotor Distance was negatively associated with Relatedness [$\beta = -1.81, SE = 0.22, p < .001$]: words with more dissimilar sensorimotor vectors were rated as less related, on average.

[12] Distributional Distance was calculated by taking the cosine distance between the final layers of BERT's contextualized embeddings for the target word in each sentence, using the bert-embedding package (https://pypi.org/project/bert-embedding/).

We then compared the Akaike Information Criterion, or AIC, of a number of different models predicting Relatedness. The models were constructed to probe the explanatory value of the distance measures, as well as the categorical condition variables (e.g., Sense Boundary). Each statistical model under consideration contained at least one of the following variables: Sensorimotor Distance (SM), BERT Distance (BERT), Sense Boundary (S), Ambiguity Type (AT), and an interaction between Sense Boundary and Ambiguity Type (S * AT).

Crucially, the inclusion of Sensorimotor Distance consistently improved model fit. In other words, the CS Norms appear to capture information that is at least partially independent from the information encoded by factors such as BERT Distance, Sense Boundary, and Ambiguity Type. Of course, it is also important to note that Sense Boundary was by far the best predictor of Relatedness, suggesting that neither distributional similarity nor sensorimotor similarity are sufficient to account for the possible effect of categorical sense representations (see Figure 9).

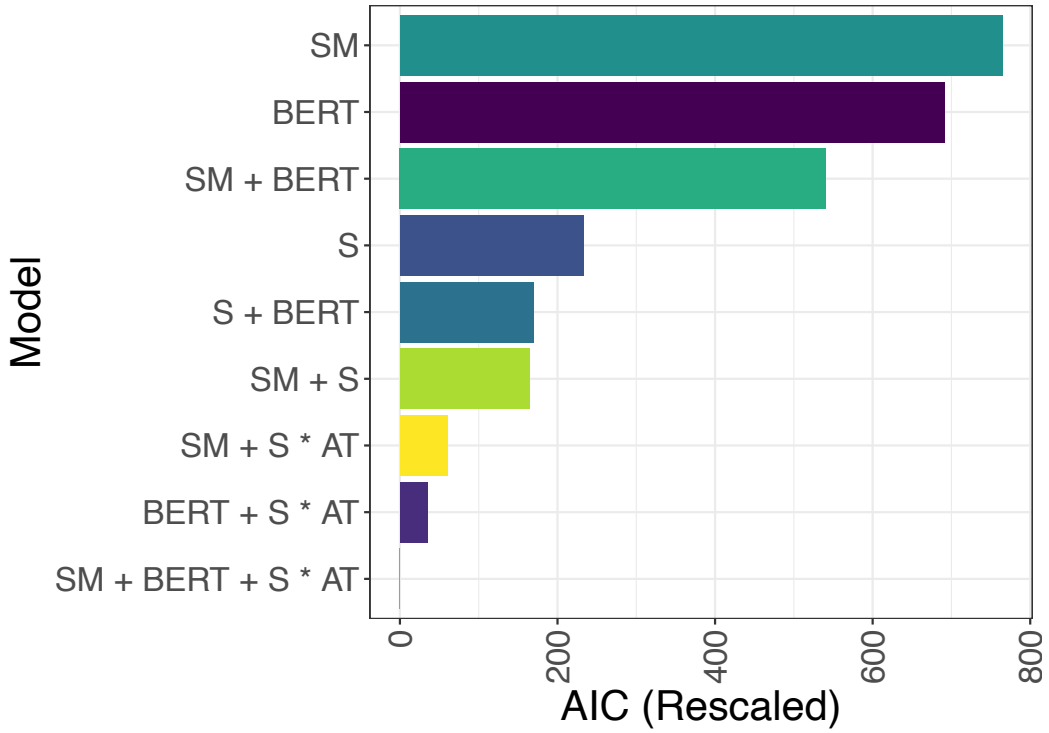


Figure 9: Rescaled AIC values for models predicting Relatedness using an assortment of factors: Sense Boundary (S), Ambiguity Type (AT), Distributional Distance (BERT), and Sensorimotor Distance (SM). A lower AIC score corresponds to better model fit.

**Comparison to Lancaster Sensorimotor Norms**

We also constructed a "baseline" model using the decontextualized LS Norms. We first identified the disambiguating word across each pair of sentence contexts (e.g., "*furry* bat" vs. "*wooden* bat"); then, we calculated the cosine distance between the decontextualized LS Norms corresponding to each of those words (e.g., between "furry" and "wooden"). We called this measure the Decontextualized Distance. Because not every disambiguating word was included in the LS Norms (e.g., "double-sided"), we excluded some pairs from the analysis, resulting in 576 pairs total.

Crucially, a linear mixed effects model including both Contextualized Sensorimotor Distance and Decontextualized Distance explained more variance than a model omitting only Contextualized Sensorimotor Distance [$\chi^2(1) = 142.07, p < .001$]. This indicates that the contextualized measure (i.e., from the CS Norms) encodes additional information beyond what can be inferred by simply identifying the sensorimotor properties of other words in the context.

## Discussion

Embodied experience appears to be crucial for how humans learn and understand language (Bergen, 2012; Pulvermüller, 1999; Barsalou, 1999), yet most large language models (LMs) are exposed to linguistic input alone (Bender and Koller, 2020). One solution is to *augment* and *evaluate* LM representations using psycholinguistic resources, such as human judgments of the sensorimotor features associated with a word (Lynott et al., 2019). However, this approach must also contend with the challenge of lexical ambiguity. Words mean different things in different contexts (Rodd et al., 2004; Trott et al., 2020), yet many lexical resources collect judgments about words in isolation.

We attempted to address this challenge by collecting judgments about the salience of various sensory modalities (e.g., *Vision*) and action effectors (e.g., *Torso*) for the same English word, in distinct sentential contexts (e.g., "flea *market*" vs. "housing *market*"). We called this dataset the Contextualized Sensorimotor Norms (CS Norms).

These contextualized norms capture variance in sensorimotor associations beyond the information already provided by the Lancaster Sensorimotor Norms. We also replicated past work (Gilhooly and Logie, 1980) suggesting that the psychological dominance of a meaning is correlated with its sensorimotor strength. Third, we found that the sensorimotor distance between contexts of use was correlated with the existence of sense boundary. We also

demonstrated the predictive utility of the CS Norms above and beyond large LMs such as BERT and the decontextualized LS Norms.

Beyond its use for NLP applications, the CS Norms could also be used to address questions of theoretical interest, such as the relative contribution of different sources of information (e.g., distributional vs. sensorimotor associations) to semantic representations in the mental lexicon (Andrews et al., 2014; Davis and Yee, 2021), as well as a role for discrete, symbolic representations (e.g., sense boundaries). For example, as with the LS Norms, researchers could use sentences from this dataset as stimuli in behavioral or neuroscientific experiments.

**Limitations**

This dataset is not without limitations.

First, it is restricted in size and breadth: 448 sentences (112 words, with 4 sentences each), in English only. In contrast, the Lancaster Sensorimotor Norms contain judgments of almost 40,000 English words (Lynott et al., 2019), and have now been extended to French (Miceli et al., 2021), Dutch (Speed and Brybaert, 2021), and more. Having demonstrated the utility of the CS Norms on a small subset of English words, one obvious direction for future research would be to expand this dataset—including more words, more senses and sentences per word, a wider variety of sentences (i.e., both experimentally controlled and naturalistic sentences), and additional languages. Similarly, existing datasets on lexical ambiguity (Haber and Poesio, 2021; Karidi et al., 2021; Schlechtweg et al., 2021; Erk et al., 2013) could be augmented with sensorimotor judgments. Further, because the original RAW-C items were adapted from psycholinguistic studies (Trott and Bergen, 2021), those items might be skewed

towards the phenomena those researchers were interested in; for example, it is possible that certain polysemous relationships (metaphor and metonymy) may be overrepresented.

A second, related problem is that the participants were sourced from an undergraduate population, which is likely non-representative of the broader population at large (Henrich et al., 2010). Similarly, the sentences themselves were hand-crafted, and thus do not reflect the full diversity of contexts these meanings might enjoy in naturalistic usage. Future work should attempt to ensure diversity in both the sample of annotators and the sentences under consideration.

Third, as others have noted (Bender and Koller, 2020; Bisk et al., 2020; Tamari et al., 2020; Borghi et al., 2019), *grounding* goes beyond sensorimotor associations. Linguistic meaning is also grounded in social experience and interaction. Recent work has attempted to incorporate these social aspects of grounding, either by integrating social information into distributional models (Johns, 2021) or simply by including more dimensions in the grounded feature representations (Binder et al., 2016).

Finally, recent work has enjoyed some success in learning grounded feature vectors directly from LM representations, typically for words rated in isolation (Turton et al., 2020; Chersoni et al., 2020; Utsumi, 2020). One question is whether contextualized embeddings, derived from a large LM such as BERT, are sensitive enough to capture the fine-grained distinctions that the CS Norms encode across sentential contexts for the same word. Of course, it is possible that the CS Norms dataset is simply too small to successfully augment a LM like BERT. However, the norms could also be used as a "challenge set", i.e., to *evaluate* how much information about sensorimotor properties of a word are in principle derivable from an LM's

representations. For example, the performance of an ungrounded model like BERT could be compared to recent multi-modal models (Zellers et al., 2021a,b).

## Conclusion

We have presented a novel resource: human judgments about the strength or salience of various sensorimotor features for 112 English words, each appearing in four distinct sentential contexts. This resource was extended from past work (Trott and Bergen, 2021), and thus also contains information about the relatedness *between* sentential contexts for the same word. We provided several demonstrations of the dataset's utility, above and beyond judgments of these words in isolation (Lynott et al., 2019), as well as large LMs such as BERT.


## Acknowledgments

Chapter 3, in full, has been submitted to the 29th International Conference on Computational Linguistics. Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

# CHAPTER 4: ARE WORD MEANINGS CATEGORICAL OR CONTINUOUS?

Words mean different things in different contexts. In some cases (approximately 7% of words in English, for instance—Rodd at al., 2004), the same sequence of characters or sounds can denote meanings that appear entirely unrelated (e.g., "river *bank*" vs. "financial *bank*"). This phenomenon is typically called *homonymy* (Valera, 2020). Far more frequent (about 84% of English words, per Rodd et al., 2004) is *polysemy*—in which related meanings (e.g., "pet *chicken*" vs. "roast *chicken*") are interpreted as corresponding to different senses of a single word. In the limit, *all* words arguably have meanings that depend on context to some extent, even if not considered outright ambiguous (Hoffman et al, 2013; Elman, 2004; Yee & Thompson-Schill, 2016). For example, the word "runs" evokes subtly different actions in "the boy *runs*" and "the cheetah *runs*" (Elman, 2004); similarly, comprehenders might activate different sensorimotor representations of the word "lemon" in "she cut the *lemon*" and "she juggled the *lemon*" (Yee & Thompson-Schill, 2016).

Each of these phenomena—homonymy, polysemy, and context-dependence—is pervasive across the world's languages (Dautriche, 2015; Valera, 2020). Accordingly, multiplicity of meanings has driven research across many different disciplines, including linguistics (Tuggy, 1993; Valera, 2020), cognitive science and psycholinguistics (Rodd et al, 2004; Elman, 2004), lexicography (Krishnamurthy & Nicholls, 2000), Natural Language Processing (Navigli, 2009; Kilgarriff, 2007; Schneider et al, 2015; Karidi et al, 2021), and legal studies (Schane, 2002), to name just a few. Knowing what the range of meanings is for any given word, or the different patterns that meaning-varying words in general display, is crucial for theories of language knowledge, use, and acquisition.

Yet despite widespread interest, there remains considerable disagreement about exactly how humans represent the multiplicity of word meanings. On some accounts, humans store different lexical representations for wordforms with unrelated meanings (i.e., homonyms), but not for wordforms with multiple, related senses (i.e., polysemes) (e.g., Cruse, 1986); other accounts argue that humans maintain distinct representations for both homonyms and polysemes (Kempson, 1977). And still others eschew the notion of discrete lexical representations altogether, arguing instead that word meanings are best characterized as occupying a continuous, context-sensitive state-space (Elman, 2004; Elman, 2009). Importantly, these different accounts also echo more general issues in Cognitive Science. To what extent is human semantic knowledge constituted by discrete, symbolic representations vs. gradient, sub-symbolic systems (Miikkulainen & Elman, 1993)? Are concepts organized by their prototypes or exemplars (Malt, 1989)? Notably, while there have been attempts to adjudicate between a subset of these accounts, none of them is entirely consistent with current empirical evidence, and none can be strictly disconfirmed.

In the sections below, we first describe the testable predictions made by each of these competing accounts, as well as their theoretical limitations. We also introduce and elaborate on two novel "hybrid" accounts, which reconcile discrete sense representations with a continuous view of meaning, and which are designed to overcome the limitations of existing theories. We then report on two behavioral experiments able to adjudicate among them, paired with an analytical approach that relies on recent advances in neural language models (Devlin et al, 2018). The results are best explained by a hybrid account that allows for effects of both continuous (i.e., distance in state-space) and categorical (i.e., sense boundaries) factors. Finally, we compare the predictive power of several computational models of the novel hybrid accounts.

## The Mental Dictionary Framework

Many accounts of how word meanings are stored and represented can be grouped under the broader umbrella of the Mental Dictionary Framework. Under this view, the mental lexicon is conceptualized as a dictionary, held in long-term memory (Pinker, 1997; Elman, 2004). Each wordform maps onto a *lexical entry*, which contains information about the word's basic semantic and syntactic properties. Accordingly, ambiguous wordforms (like homonyms) map onto multiple, distinct entries, as they would in a literal dictionary. Critically, the categorical boundary between distinct word meanings is theorized to exert an influence on psychological processing above and beyond the context-dependent nature of word meaning. Put another way: there is a qualitative distinction between outright ambiguity (e.g., "delicious *port*" vs. "windy *port*") and mere under-specification (e.g., "*big* building" vs. "*big* ant").

Within this Mental Dictionary Framework, there are at least two dominant theoretical accounts. The primary distinction between these accounts is in how they treat polysemy—i.e., words with multiple, related meanings—namely, whether polysemous meanings are represented differently from homonymous meanings. According to Sense Enumeration Accounts, polysemy is represented much like homonymy: all ambiguous words map onto multiple, distinct lexical entries (Kempson, 1977). That is, just as "financial *bank*" and "river *bank*" would constitute distinct entries, so too would "pet *chicken*" and "roast *chicken*". Sense Enumeration Accounts are considered by some (Klepousniotou, 2002; Pustejovsky, 1995) to be uneconomical; because polysemy is extremely pervasive (Rodd et al, 2004), storing each polysemous meaning separately results in a proliferation of lexical entries. Nonetheless, the chief advantage of Sense

Enumeration Accounts is that they sidestep the difficulty of addressing irregular forms of polysemy, i.e., cases in which multiple meanings are related but not in a systemic fashion (Kempson, 1977; Rice, 1992). Sense Enumeration Accounts make two concrete predictions about cognitive processing. First, pairs of related senses (e.g., "pet *chicken*" vs. "roast *chicken*") should be distinguishable in behavior from pairs of same sense uses (e.g., "roast *chicken*" vs. "marinated *chicken*"). And second, because polysemy is represented in the same fashion as homonymy, the behavior of words with what are classified as related senses should *not* be distinguishable from those with homonymous senses.

Core Representation Accounts also view the lexicon as storing discrete entries. But unlike Sense Enumeration Accounts, they do not view multiple related senses as separate lexical entries, instead deriving or generating meanings during online language processing from a single "core" representation (Cruse, 1986; Pustejovsky, 1995; Pustejovsky & Bouillon, 1995; Pustejovsky, 2002). For Core Representation Accounts, the mental lexicon contains not only lexical entries, but also *rules*—much like grammar—for systematically extending word senses as a function of context. The generative lexicon (Pustejovsky, 1995) is one well-known example of a Core Representation Account; Pustejovsky (2002) motivates this additional component by appealing both to parsimony and the underlying systematicity by which meanings are extended. In a generative lexicon, lexical entries are associated with some minimal semantic configuration––what Pustejovsky calls their *qualia structure*—which affords (or precludes) particular inferences when composed with other lexical items. For instance, the wordform "bake" would unambiguously denote a *change-of-state* process, but the interpretation of this process as *change-of-state* or *creation* would be constrained by the speaker's choice of direct object (e.g., "potato" vs. "cake"). This in turn places constraints on the interpretation of the verb. Core Representation

Accounts thus make distinct predictions from Sense Enumeration Accounts. Most notably, because polysemous meanings are represented in a different fashion from homonymous meanings, polysemy and homonymy should elicit measurably distinct behavior in comprehenders. Further, under a stronger interpretation[13] of Core Representation Accounts, same sense uses of a word (e.g., "wrapping *paper*" and "shredded *paper*") should not exhibit enhanced facilitation (e.g., in priming, memory, etc.) above and beyond a neutral baseline (e.g., "____ *paper*"); since the shared core is activated each time the wordform is encountered, even in the baseline condition, the target meaning should be equally accessible (Klein & Murphy, 2001; Klein & Murphy, 2002).

Experimental research offers mixed evidence on these accounts. On the one hand, polysemy does appear to elicit distinct behavior from homonymy. In lexical decision tasks, words categorized as homonymous are recognized more slowly than those categorized as polysemous (Rodd et al, 2002; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Armstrong & Plaut, 2008), possibly because the unrelated meanings compete during lexical access (Rodd et al, 2002). Homonymy may also be more challenging to learn than polysemy (Rodd et al, 2012; Floyd & Goldberg, 2021). These findings are inconsistent with predictions of the Sense Enumeration Account, and in turn favor the Core Representation Account.

On the other hand, senses ostensibly related through polysemy elicit distinct behavior from same sense uses of a wordform (Klein & Murphy, 2001; Klein & Murphy, 2002; Yurchenko et al, 2020). In a memory task (Klein & Murphy, 2001), subjects were worse at recognizing previously observed wordforms (e.g., "wrapping *paper*") when the repeated phrase

[13] As Klein & Murphy (2001) note, it is technically possible to reconcile this result with a Core Representation Account: if accessing specific meanings (e.g., "wrapping *paper*") requires the application of a generative "rule", and if the core representation is under-specified, it might be easier to transition between same sense meanings (because the generative rule is already activated) than between an under-specified core and a specific meaning.

employed them in a context evoking a polysemously related sense (e.g., "liberal *paper*") than a same sense context (e.g., "shredded *paper*"). Similarly, in a primed sensibility judgment task, subjects were also less accurate when responding to polysemously related senses than same sense uses or a neutral baseline (e.g., "____ *paper*") (Klein & Murphy, 2001), and displayed differentiable brain responses in an EEG experiment (Yurchenko et al, 2020). These findings are sometimes interpreted as evidence against the Core Representation Account (Klein & Murphy, 2001). Indeed, they do disconfirm a strong view in which comprehenders process and represent related senses identically to same sense uses.

Importantly, however, given that most Core Representation Accounts argue that related senses are derived via a generative rule, it is technically possible to reconcile these accounts with the finding that related senses are processed differently from same sense uses, since the application of this rule might increase processing time (Klein & Murphy, 2001). Some Core Representation Accounts can also accommodate the difference in facilitation between same sense primes and a neutral baseline—if generative rules can be primed, then it should be easier to transition between same sense meanings (given that the rule is already primed) than between an under-specified core representation and a more specific meaning (which would require activating the rule for the first time). Thus, these particular results do not allow us to distinguish between Sense Enumeration Accounts and a more nuanced version of Core Representation Accounts.

This leaves considerable uncertainty. Polysemous meanings may indeed be represented separately (as in Sense Enumeration Accounts), or at least enjoy some degree of functional separation (as more nuanced versions of both accounts would predict); but according to some evidence, this representational mechanism appears to be distinct from homonymy (as in Core Representation Accounts). And in fact, recent work suggests that the simple distinction between

polysemy and homonymy (and even between same and different senses) may be overly simplistic. Multiple studies (Klepousniotou, 2002; Klepousniotou & Baum, 2007; Bambini et al, 2013; Lopukhina et al, 2018; Yurchenko et al, 2020) have found differences in behavioral and neurophysiological responses to pairs of meanings related via different polysemy mechanisms: *metonymy* (e.g., "pet *chicken*" vs. "roast *chicken*") and *metaphor* (e.g., "polluted *atmosphere*" vs. "relaxed *atmosphere*"). Similarly, other studies (Klepousniotou et al, 2008; Brown, 2008) have found that measures of processing ease (e.g., accuracy and response time) are predicted by the *degree of overlap* or *semantic similarity* between two senses.

Importantly, current evidence as described above does not fully adjudicate between the two accounts falling under the Mental Dictionary Framework. Moreover, other approaches identify certain limitations of this framework and attempt to address them.

## Challenges to the Mental Dictionary Framework

The Mental Dictionary Framework—at least as outlined above—has been challenged on several theoretical grounds. Some of these arguments relate specifically to the question of lexical ambiguity, while others concern the role of knowledge and context more generally (Elman, 2004).

### Identifying sense boundaries is challenging.

The Mental Dictionary Framework reifies the lexicographic concept of discrete word senses, which requires a commitment as to whether the difference in meaning conveyed by a given pair of word uses corresponds to ambiguity (i.e., distinct senses) or mere context-dependence (sometimes called *vagueness*).

This distinction may appear obvious in some cases (e.g., "river *bank*" vs. "financial *bank*" are readily interpreted as distinct senses), but in many situations, it is difficult to pin down using standard linguistic tests (Tuggy, 1993; Geeraerts, 1993; Hanks, 2000; Kilgarriff, 2007; Krishnamurthy & Nicholls, 2000). Tuggy (1993) illustrates the challenge using the verb "paint", which can describe a number of conceptually related actions, including: 1) painting a portrait in oils; 2) painting a landscape with watercolors; 3) painting stripes on the parking lot; 4) applying makeup to the face; and more. (1) and (2) plausibly belong to the same sense, but (1) and (3) may seem anomalous when used in *zeugmatic cross-reference* ("I'm painting [a portrait] and Ben is painting [stripes on the road] too"), a common test for distinguishing ambiguity from vagueness. According to that criterion, then, (1) and (3) should be considered distinct senses, suggesting that "paint" is at least partially ambiguous. Yet even this standard test is not without limitations. First, other contexts may permit a crossed reading of these two meanings ("When I'm painting [a portrait], I try to get the color on evenly, and so does Jane when she paints [stripes on the road]") (Tuggy, 1993). Second, it is not always clear whether the anomalous reading arises from lexical ambiguity per se. For example, the sentence "the newspaper costs $0.50 and sacked all its staff" seems anomalous, but is that because human comprehenders represent two distinct senses for the wordform "newspaper", or because of difficulties that arise during a more general-purpose process of pragmatic interpretation (Kilgarriff, 2007)? In other cases, cross-reference can elicit zeugma in the absence of ambiguity; for example, "I evicted and knew her" does not seem to involve lexical ambiguity, but many readers likely find it anomalous (Kilgarriff, 2007).

Further, as others have noted (Geeraerts, 1993; Kearns, 2006), different tests are often in conflict with one another. Two uses of "book" might have different truth conditions and

accompany distinct modifiers (e.g., a cultural artifact vs. a physical object), but may still permit cross-reference ("I'm enjoying [this book] but I wish it had larger print") (Kearns, 2006, pg. 369-370).

Of course, the fact that ambiguity is sometimes hard to distinguish from context-dependence is not evidence that the distinction itself is in principle invalid. Perhaps we simply need better tests, or the existing tests ought to be weighted in more sophisticated ways. But it does indicate that the situation is more complex than it might appear at first glance, particularly when we apply this distinction to the mental representation of word meaning—and surprisingly, there is a dearth of studies investigating the psychological reality of discrete word senses (as distinct from context-dependence) in the first place.

**Homonymy and polysemy (and context-dependence) are not easily distinguished.**

The distinction between homonymy and polysemy is also notoriously challenging to define and detect (Tuggy, 1993; Valera, 2020). These forms of lexical ambiguity are typically distinguished in one of two ways: a) determining whether a given pair of meanings shares a common *etymon* or etymological source (polysemy) or not (homonymy); and b) determining whether a given pair of meanings is conceptually similar or related (polysemy), or not (homonymy) (Valera, 2020). Yet both methods have limitations. Shared etymology is difficult to establish and does not entail synchronic psychological association—even if two meanings were once related, the phenomenon of *semantic drift* can lead to those meanings drifting apart over time, leading to apparent homophony (Tuggy, 1993). For example, the words *flour/flower* actually began as a borrowing from the same etymon (*flur*) from French (*fleur*, meaning both "blossom" and "the choicest part of something"); these meanings drifted in various ways (e.g., "flower of wheat" referring to the

endosperm of wheat), and in fact retained the same spelling until the 18th century—now the words are heterographic homonyms (Jurafsky, 2014).

Psychological relatedness seems preferable in principle if our goal is to establish theories about the structure of the mental lexicon. But assessing psychological relatedness raises thorny definitional and methodological questions: how exactly is "relatedness" established? Should some mechanisms or manners of conceptual relation—such as metaphor or metonymy—be weighted more heavily than others, or does any manner of relation count? Moreover, as others have noted (Klepousniotou, 2002; Valera, 2020) the very notion of relatedness—and the way it's usually measured—lies on a *continuum*, as opposed to a dichotomy. This has led some (e.g., Deane, 1988; Tuggy, 1993) to suggest that homonymy, polysemy, and under-specification ought to be considered as lying along a cline as well:

> "In effect, the three types form a gradient between total semantic identity and total semantic distinctness" (Deane, 1988, pg. 327).

> "Ambiguity and vagueness may be seen as occupying opposite ends of a continuum with polysemy in the middle." (Tuggy, 1993, pg. 1).

This continuum view is to some extent compatible with current psycholinguistic evidence. As noted earlier, processing ease (as measured by RT, accuracy, N400 effects, etc.) on several different tasks (e.g., lexical decision, primed sensibility judgments, etc.) is predicted not only by the coarse distinction between homonymy and polysemy, but also by the *degree of overlap* between two meanings (Klein & Murphy, 2002; Klepousniotou et al, 2008; Brown, 2008).

However, a cline between ambiguity and under-specification does not square easily with the Mental Dictionary Framework. Under a strong interpretation of "continuum", categories such as homonymy and polysemy are helpful descriptive abstractions, but are not viewed as psychologically real; it is challenging to reconcile this position with the Mental Dictionary Framework, in which word meanings are represented in discrete entries. This suggests that an alternative framework might be required—one that allows for greater flexibility of representation and context-dependence.

**Word meaning is flexible and context-dependent.**

A more general critique of the Mental Dictionary Framework is presented by Elman (2009), who argues that in general it cannot adequately address the dynamic, context-dependent nature of word meaning. Elman (2009) reviews a large body of psycholinguistic research, demonstrating that words encode detailed world knowledge, and that this knowledge appears to play an early role in sentence processing. This includes early detection of incompatible or unlikely instrument/patient pairings (e.g., *Susan used the scissors to cut the expensive wood*), the ability of discourse context to override typical verb/patient pairings (e.g., a "shopping" context renders *the lifeguard saved* ***money*** easier to process, even though the default expectation might be *saved* ***lives***), and more. In other words, "lexical representations contain a significant amount of detailed word-specific information that is available and used during online sentence processing" (Elman, 2009, pg. 566).

For Elman, this raises the question of *which* information is included in these lexical representations. Overly sparse entries (e.g., a phonological representation and part of speech) cannot account for the early effects of lexical knowledge; but if we instead add sufficient detail

to these entries to accommodate the psycholinguistic evidence, it results in a combinatorial explosion (e.g., storing all the possible instrument/patient contingencies and their respective compatibilities). Elman (2004; 2009; 2011) ends up rejecting the notion of discrete lexical entries altogether, instead advocating for a view in which word meaning is represented as *trajectories* through a continuous state-space. This alternative view, which we call the Continuity of Meaning Framework, is described in more detail below.

## The Continuity of Meaning Framework

In the Continuity of Meaning Framework, words are conceptualized as *cues to meaning*––eliciting context-dependent trajectories through a continuous state-space, as in a recurrent neural network (Elman, 2004; Elman, 2009; Elman 2011; Li & Joanisse, 2021). In theory, the dimensions of this state-space could be constituted by many different features of lexical experience, including the distributional statistics or usage patterns of a word (Li & Joanisse, 2021), as well as sensorimotor associations with that word (Elman, 2011). In this paper, we focus primarily on the role of distributional patterns, but a potential role for sensorimotor correlates is considered in the General Discussion.

The precise trajectory elicited by a particular word token (e.g., "runs") will necessarily be contingent on the prior state of the network, which in turn is entirely dependent on context (e.g., "the boy *runs*" vs. "the cheetah *runs*"). Thus, this approach builds the role of context directly into its conception of word meaning: rather than positing discrete *senses* for two contexts of use, the difference in meaning can be captured by the different trajectories elicited by "runs" across those two sentential contexts. Accordingly, when the same wordform is encountered in contexts that differ to a greater degree, it will also elicit trajectories through the network that differ more–

–the distance in state-space between "the boy *runs*" and "the cheetah *runs*" should be smaller than the distance between "the boy *runs*" and "the clock *runs*". This yields the theoretical benefits of word "types" without the disadvantages discussed above (Elman, 2004), in the following way. To the extent that two tokens elicit similar trajectories in state-space, they behave quantitatively like a common "type" of sorts—but while also differing in subtle, context-dependent ways. This framework also reflects a larger paradigm shift towards continuous accounts of cognitive processes more generally (Spivey & Dale, 2004; Spivey & Dale, 2006; Spivey, 2008); increasingly, many processes thought to consist of discrete operations carried out over symbolic representations have been modeled using a dynamical systems approach that posits no explicit representations (Spivey, 2008; Beer, 2003; Chemero, 2011).

How might this framework handle the problem of lexical ambiguity? In its strongest theoretical implementation, the notion of discrete sense categories is rejected altogether. This view—which we call Pure Exemplar Theory—holds that discrete meaning categories for a word (i.e., "senses") is a convenient theoretical abstraction, but is not psychologically real. A "sense" is simply a label describing a stable pattern of activity within the high-dimensional state-space. According to Pure Exemplar Theory, the difference between lexical ambiguity and context-dependence is entirely a matter of degree: all words elicit variable trajectories through state-space, and although we might decide that some of these trajectories are better described in terms of multiple "sense-clusters", this distinction is not assumed to be cognitively relevant—it does not influence cognitive processing above and beyond the *distance* in state-space between any two contexts of use. This theory thus has an affinity to other accounts of language processing that eschew stored abstractions (Ambridge, 2020).

As a consequence, on Pure Exemplar Theory, the difference between homonymy and polysemy is also one of degree, not kind. Homonymy corresponds to words with more distant, differentiable contexts of use, while polysemy corresponds to words whose contexts of use are closer in state-space. A similar account is presented in Rodd (2020), in which these phenomena are understood from the perspective of attractor basins. Homonymous meanings correspond to distant, deep attractor basins, while polysemous meanings correspond to shallow, more connected basins.[14]

To date, Pure Exemplar Theory cannot be strictly disconfirmed by any existing psycholinguistic research. Merely finding a difference in how comprehenders process homonymous or polysemous words, as many studies have (Rodd et al, 2002; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Armstrong & Plaut, 2008), does not rule out the possibility that this difference simply reflects distances in a fundamentally continuous space; if homonymous meanings are more distant than polysemous meanings, then Pure Exemplar Theory predicts that they should be harder to process. The same goes for finding a difference between how people process ostensibly same sense and different sense uses (Klein & Murphy, 2001): if same sense uses are used in more similar contexts, then Pure Exemplar Theory predicts that they should be easier to process than different sense uses. Thus, Pure Exemplar Theory currently offers a better explanation of existing data than either theory falling under the Mental Dictionary Framework, as it accommodates findings distinguishing homonymy and polysemy, as well as ambiguity from context-dependence. It is also more consistent with other evidence that is harder to reconcile with either account, such as the finding that the degree of overlap between two

[14] Rodd (2020) does not necessarily argue for some form of the Pure Exemplar Theory. Rather, it is the closest example of a model of lexical ambiguity in which meaning is seen as distributed feature-vectors in a continuous landscape. It is possible that Rodd's (2020) state-space model is compatible with a cognitive distinction between ambiguity and context-dependence.

meanings—or more aptly, between two *contexts of use*—influences behavior (Klein & Murphy, 2002; Klepousniotou et al, 2008; Brown, 2008).

## Hybrid Meaning Framework: Category Effects in a Continuous State-Space

Although Pure Exemplar Theory cannot be disconfirmed by current empirical evidence, there are at least two reasons to think it might not stand up to a more targeted falsification attempt.

First, outside of lexical ambiguity, there are a number of domains in which humans treat continuously varying input as falling into discrete categories that have psychological effects above and beyond continuous variation in that input (Goldstone & Hendrickson, 2010). This phenomenon, categorical perception, transforms the perceptual space "such that differences between objects that belong in different categories are accentuated, and differences between objects that fall into the same category are deemphasized" (Goldstone & Hendrickson, 2010, pg. 69). Evidence for categorical perception is often demonstrated by manipulating a continuous stimulus, such as voice onset time or color hue, and asking whether behavioral or neurophysiological responses to that stimulus exhibit discontinuity. Many of these domains involve language in some way (though not all, e.g., face perception). For example, responses to continuous variation in acoustic input exhibits discontinuity dependent on the phoneme categories of a language (Liberman et al, 1957). Similarly, neurophysiological responses to variation in color hue are dependent on language-specific color categories (Thierry et al, 2009; Mo et al, 2011). This phenomenon also extends to objects, i.e., whether two distinct referents are co-categorized by the lexicon of a language. English speakers distinguish *cups* from *mugs*, while Spanish speakers refer to both as *taza*. Accordingly, English speakers exhibit a sharper visual

mismatch negativity effect when viewing pictures of mugs interspersed with those of cups (or vice versa), than do Spanish speakers (Boutonnet et al, 2013). While this last example involves distinct referents (i.e., cups and mugs) rather than continuous variation in a perceptual stimulus (e.g., color hue), it remains relevant to the question of lexical ambiguity. If sense categories are psychologically real, one might expect them to exert a similar influence: that is, the conceptual distance between two contexts of use should be magnified if those contexts straddle a sense boundary—and compressed if they fall within a single sense category.

Second, recent empirical evidence from an offline task (Trott & Bergen, 2021) is broadly consistent with this prediction. Trott & Bergen (2021) asked participants to rate the conceptual relatedness of the same wordform in two different contexts of use. In some cases, these contexts corresponded to the same sense (e.g., "marinated lamb" vs. "roasted lamb"), while others corresponded to different senses (e.g., "marinated lamb" vs. "friendly lamb"). Additionally, some different-sense pairs were classified (according to dictionaries) as polysemous (e.g., "marinated lamb" vs. "friendly lamb"), while others were homonymous (e.g., "furry bat" vs. "baseball bat"). Participants' ratings were compared with a continuous measure of the distance between these contexts of use, obtained using the neural language model BERT[15] (Devlin et al, 2019). As expected, more distant contexts were rated as less related (*Pearson's r = 0.58*). Critically, however, BERT consistently *underestimated* how related participants found same-sense pairs to be, and *overestimated* how related they found different-sense homonyms to be (Trott & Bergen, 2021). Both results point to the possibility that participants' relatedness judgments were influenced not only by continuous variation across contexts of use, but also by human sense categories. According to this interpretation, sense categories compressed the

[15] BERT is described in more detail in the Current Work section.

conceptual distance between same-sense pairs and amplified the distance between distance-sense pairs (particularly for homonyms).

On the other hand, there are several important limitations to this result. First, as participants' relatedness judgments were made offline, it remains unclear whether putative sense categories play an active role in shaping online word processing in context. Second, participants were explicitly asked to rate meaning similarity on a labeled scale from 1 ("totally unrelated") to 5 ("same meaning"). This might have encouraged participants to draw on meta-linguistic category knowledge to complete the task, even if such knowledge does not actually influence the course of "ordinary" language comprehension. Together, these limitations imply that we cannot yet rule out the Pure Exemplar Theory as a viable account of the mental lexicon.

Of course, as noted in the previous section, there are also a number of limitations to both theories falling under the Mental Dictionary Framework. This raises the possibility of a hybrid account—one that reconciles the notion of discrete sense categories with a continuous, graded meaning space.

**Hybrid Meaning Theory.** Hybrid Meaning Theory posits the existence of senses (or "sense-clusters"). These sense categories warp the underlying continuous context space according to which category a particular point or trajectory within that space belongs to. Specifically, contexts of use belonging to the *same* sense category should become closer together, while contexts of use belonging to *different* sense categories should become further apart.

Importantly, this theory requires that the co-categorization of two contexts of use depend on some factor other than distance in context space. That is, Hybrid Meaning Theory is not merely an exaggeration of existing clumpiness. Rather, it requires that contexts of use are

somehow *assigned* to distinct sense categories, which themselves are derived from a source of information or representation external to that context space—and which, in turn, warp the distance between those usage contexts. There are many possible mechanisms by which sense categories might form, including: identification of distinct sensorimotor associations for different contexts of use, distinct communicative or pragmatic contexts, and more (see the General Discussion for a more detailed description). Importantly, the primary commitment of Hybrid Meaning Theory is not to a specific categorization mechanism, but to the claim that sense categories impinge on a continuous meaning-space and transform that space in some way.

*Figure 10* presents one possible implementation of these transformations: within a sense-cluster, points attract towards the centroid of that sense category, resulting in an exaggeration of conceptual distance across clusters. We call this mechanism the Sense Attraction account.

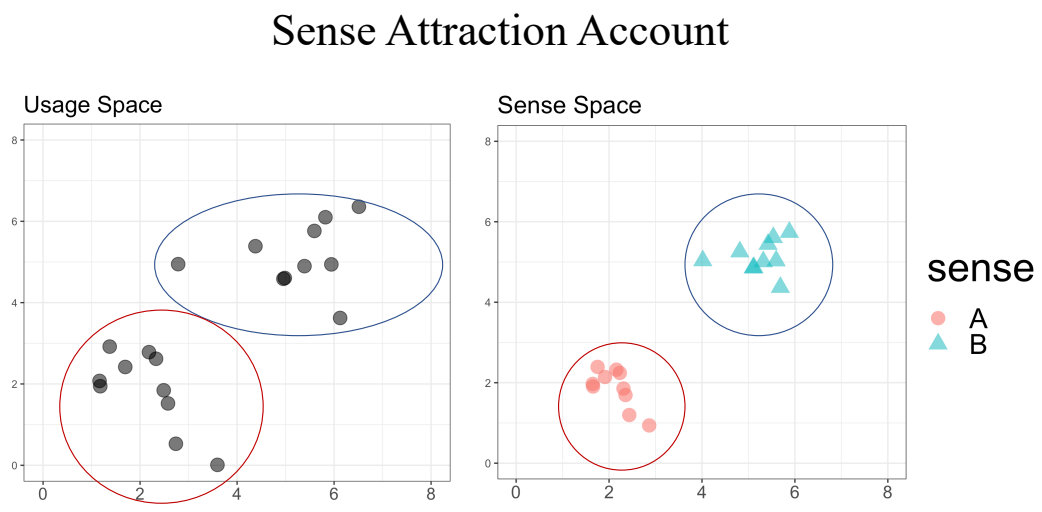


Figure 10: In the Sense Attraction Account, existing clumpiness in usage-space is exaggerated. For within-cluster uses of a wordform, contextual distance is compressed in meaning-space; for across-cluster uses of a wordform, contextual distance is amplified.

Another possible mechanism, which we call the Sense Distillation Account, is illustrated in *Figure 11*. Unlike the Sense Attraction Account, within-cluster variance is *distilled* into a single point, i.e., the centroid of that cluster. Critically, this preserves the metric properties of the continuous space: clusters with centroids that are relatively closer together will result in sense representations that are also closer in meaning-space. But because within-cluster variance is removed, the Sense Distillation Account predicts that the difficulty of transitioning between two within-sense contexts of use is not predicted by their distance in usage-space—whereas the Sense Attraction Account predicts that within-cluster variance should predict processing difficulty even for same sense uses of a wordform.

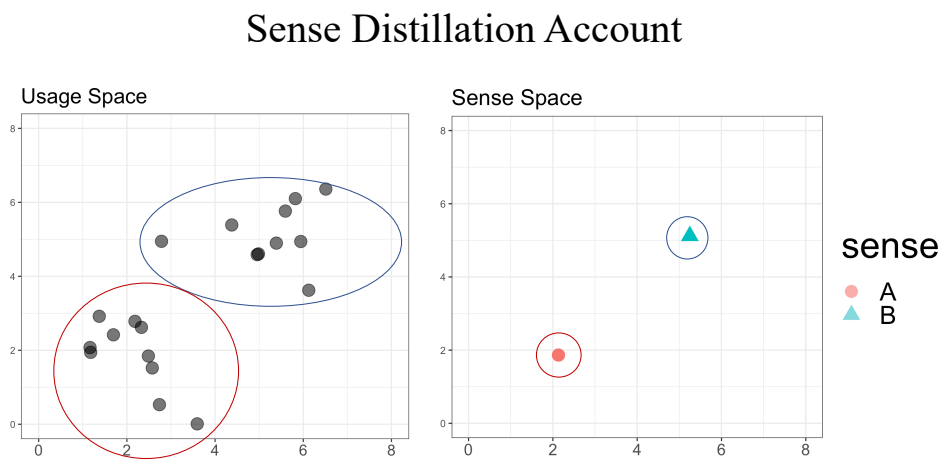


Figure 11: In the Sense Distillation Account, clusters are distilled into their centroids. This removes within-cluster (i.e., within-sense) variance entirely, but preserves the underlying metric properties of the continuous space.

Both possible accounts outlined above are analogous to more general cognitive mechanisms implicated in the resolution of continuously varying or ambiguous input into

discrete categories, such as categorical perception (Goldstone & Hendrickson, 2010) or the Ganong Effect (Ganong, 1980). As described earlier, these theoretical mechanisms have played an important role in accounting for human behavior in other domains (e.g., speech perception); we propose that an analogous mechanism could be of use in explaining human lexical knowledge.

Both implementations of Hybrid Meaning Theory (Sense Compression and Sense Distillation) also acknowledge the importance of continuity and context-dependence, as well as the possibility that the mind carves further structure into this continuous space. While this theory has not been directly tested, its stipulation of continuous gradation in meaning allows it to accommodate existing evidence for the dynamic, flexible nature of word meaning (Elman, 2009). Further, its representation of category structure makes it consistent with evidence that discrete sense representations play a role in cognitive processing (Klein & Murphy, 2001; Klein & Murphy, 200; Yurchenko et al, 2020). It also makes a concrete prediction that differentiates it from Pure Exemplar Theory, as well as from both theories falling under the Mental Dictionary Framework. Namely, the ease of transitioning between two contexts of use, as in primed sensibility judgment tasks (Klein & Murphy, 2001; Klepousniotou et al, 2008; Brown, 2008) should be affected both by the distance in usage space between those contexts *and* whether or not a sense boundary separates those uses.

Nevertheless, at present, there is no reason to prefer this theory over the more parsimonious Pure Exemplar Theory, which has not yet been disconfirmed and which also accommodates existing evidence.

**Hybrid+ Theory.** The Hybrid Meaning Theory described above claims that discrete sense categories are integrated with a continuous meaning-space. Yet there is also evidence that human sense knowledge is further shaped by the kind of ambiguity at play (Rodd et al, 2002; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Armstrong & Plaut, 2008; Trott & Bergen, 2021). If this is true, the Hybrid Meaning Theory must be augmented with a categorical distinction between homonymy and polysemy—above and beyond distances in state-space. We call this augmented view the Hybrid+ Theory.

As noted earlier, the Continuity of Meaning Framework predicts that on average, pairs of homonymous senses are likely to occupy more distant regions of sense-space than pairs of related senses (Rodd, 2020). That is, homonyms and polysemes occupy a continuum of *proximity in sense-space* ranging from very close to very distant. Neither Pure Exemplar Theory nor Hybrid Meaning Theory categorically distinguishes the two phenomena. In principle, however, it is conceivable that the human mind transforms a continuous space not only with discrete sense representations, but also in a way that reflects distinct kinds of lexical ambiguity. This transformation could have the result of differentiating homonymy and polysemy above and beyond the proximity of their sense-clusters in usage space (Rodd et al, 2002; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Trott & Bergen, 2021). We call this modified view Hybrid+ Theory, given that it posits both discrete sense representations and distinct kinds of relationships between these sense representations, all atop continuous effects of context.

As with sense categories, there are a number of reasons why a categorical difference between polysemy and homonymy could emerge. For one, various theories of lexical representation propose that they are realized through different cognitive mechanisms (Cruse, 1986), which could produce categorically distinct behavior. Additionally, polysemy is

systematic, both within and across languages (Srinivasan & Rabagaliatti, 2015), which might scaffold the learning of new polysemous meanings of known wordforms (Srinivasan & Snedeker, 2011) as compared with unrelated meanings of the same wordform (e.g., "dog *bark*" and "tree *bark*"). In theory, this differentiation could also occur along multiple levels of granularity, distinguishing not just homonymy and polysemy but also different kinds of polysemy, such as metaphor and metonymy (Yurchenko et al, 2020). Differentiation could even occur at the level of specific semantic relations, such as Animal/Meat or Material/Product (Srinivasan & Snedeker, 2011; Srinivasan & Rabagliati, 2015).

## Current Work

Thus far, we have reviewed several theories of how humans process and represent word meaning, with a particular focus on ambiguous words. The Mental Dictionary Framework views word meanings as analogous to entries in a dictionary; each unique form-meaning pairing is represented in a *lexical entry*, with ambiguous words (like homonyms) corresponding to multiple lexical entries. The Mental Dictionary Framework can be further subdivided into accounts that distinguish between polysemy and homonymy (Core Representation Accounts) and those that view all ambiguous words as mapping onto distinct lexical entries (Sense Enumeration Accounts). Crucially, both kinds of account claim that word senses are psychologically real and constitute categorical representations in the mind.

In contrast, The Continuity of Meaning Framework views word meanings as trajectories in a continuous, context-sensitive state-space. In Pure Exemplar Theory, the notion of discrete sense representations is rejected altogether, along with the categorical distinction between homonymy and polysemy (Elman, 2009).

We also described two novel, "hybrid" theories falling under the Hybrid Framework. Hybrid Meaning Theory claims that meaning is constituted by a continuous state-space, but also that existing "clumpiness" in a word's pattern of use is *exaggerated* by the mind (see *Figures 1-2*). The Hybrid+ Theory takes this model one step further, and claims that the mind further differentiates between homonymy and polysemy in this continuous space.

To test these theories, we selected a methodological paradigm—primed sensibility judgments—that has previously been used to demonstrate categorical effects of sense boundaries (Klein & Murphy, 2001; Yurchenko et al, 2020), as well as a distinction between homonymy and polysemy (Klepousniotou et al, 2008; Brown, 2008). Specifically, processing difficulty—as indexed by response time (RT) and accuracy—is increased when the uses of an ambiguous wordform across a prime and target sentence correspond to what are classified as different *senses* (Klein & Murphy, 2001; Yurchenko et al, 2020); this effect is larger for different-sense sense pairs classified as homonyms (Brown, 2008), or with less semantic overlap (Klepousniotou et al, 2008), than for words that are closely related.

Each theory makes distinct, testable predictions about which variables should influence behavior, and which should not. This means that each theory (with the exception of Hybrid+ Theory) can be disconfirmed by finding that some variable of interest (e.g., sense boundaries) predicts behavior when the theory claims that it should not. For example, Pure Exemplar Theory predicts that the ease of transitioning between two contexts of use (as measured by RT or Accuracy) should be predicted by a continuous measure of the distance between those contexts in usage-space—but not by whether those contexts of use span a sense boundary (e.g., "marinated lamb" and "friendly lamb") or belong to the same sense (e.g., "marinated lamb" and "roasted lamb"). Conversely, both varieties of the Mental Dictionary Framework predict an

effect of sense boundaries on behavior, but not a graded effect of contextual distance above and beyond this categorical effect. Hybrid Meaning Theory predicts both a graded effect of contextual distance *and* an effect of sense boundaries—but critically, the effect of sense boundaries should *not* be different across homonyms and polysemes (once contextual distance is accounted for).

Table 1: Each theory makes distinct, testable predictions about which factors should influence behavior.

| | **Mental Dictionary Framework** | | **Continuity of Meaning Framework** | **Hybrid Framework** | |
|---|---|---|---|---|---|
| | Sense Enumeration | Core Representation | Pure Exemplar Theory | Hybrid | Hybrid+ |
| Graded effects of context | -- | -- | Yes | Yes | Yes |
| Effect of sense boundaries | Yes | Yes | -- | Yes | Yes |
| Effect of sense boundary larger for homonyms than polysemes | -- | Yes | -- | -- | Yes |

In other words, Hybrid Meaning Theory (along with Pure Exemplar Theory, and both Mental Dictionary theories) does not predict an *interaction* between sense boundary and ambiguity type. Technically, this theory is compatible with a main effect of ambiguity type (i.e., an overall difference across homonymous and polysemous stimuli), given that different words and sentence frames will be used. In order to disconfirm the theory, we would need to observe an

interaction: a larger effect of sense boundaries for homonyms than polysemes, as observed for offline judgments in Trott & Bergen (2021). Only Hybrid+ Theory is compatible with this interaction effect. Accordingly, only this final theory cannot be strictly falsified, given that it predicts non-zero effects for all variables of interest. That said, certain patterns of results are nonetheless more compatible with alternative, simpler theories; for example, there is little reason to prefer Hybrid+ Theory if no graded effect of context is found, once categorical sense representations are accounted for (see *Table 1*).

Past work has focused primarily on adjudicating between the varieties of the Mental Dictionary Framework. Although a number of researchers have raised the possibility of homonymy and polysemy occupying a continuum (see Challenges to the Mental Dictionary Framework above), none have attempted to directly adjudicate between the Mental Dictionary Framework and Pure Exemplar Theory, nor test the Hybrid Theories introduced here. That's what the current experiments aimed to do.

**Measuring Continuous Contextual Distances**

A critical prerequisite for comparing these theories is operationalizing the notion of continuous distance in state-space. Such an operationalization must be both continuous and context-sensitive, so that one context of use (e.g., the word "lamb" in "marinated lamb") can be compared to another (e.g., in "friendly lamb"), e.g., by calculating the distance between these contexts.

To operationalize this notion of continuity, we used BERT (Devlin et al, 2019), a state-of-the-art neural language model (NLM). There is a growing body of literature using BERT and other NLMs as operationalizations of human lexical-semantic knowledge in general (Li & Joanisse, 2021; Trott & Bergen, 2021; Nair et al, 2020; Haber & Poesio, 2020a; Haber & Poesio,

2020b), and to test Elman's (2004; 2009) cues to meaning framework in particular (Li & Joanisse, 2021; Trott & Bergen, 2021). It is important to note that BERT (like most NLMs) is trained on linguistic input alone (Bender & Koller, 2020), and lacks access to any extra-linguistic sources of information that humans might use to represent the meanings of a word, such as sensorimotor associations. Thus, BERT reflects a particular operationalization of the Continuity of Meaning Framework: its representational space is continuous, and the topology of this continuous space is determined by statistical regularities in which words co-occur with which other words. While this operationalization has clear limitations (Bender & Koller, 2020), it is compatible with views of linguistic meaning that emphasize the role of usage (Wittgenstein, 1953), such as the *distributional semantic hypothesis* (Harris, 1954; Firth, 1957; Lenci, 2008). The distributional semantic hypothesis states that words with more similar meanings should appear in more similar contexts—and consequentially, that meaning similarity should be derivable from contextual similarity.

BERT (base) was trained on a large text corpus (>3 billion word tokens) using two objectives: 1) a masked language modeling task, in which the model must learn to predict a "masked" word in some sentential context (e.g., "I went to the [MASK] bank"); and 2) next-sentence prediction, in which the model must learn to predict whether two sentences occurred next to each other. After training, BERT can be used to produce *contextualized embeddings* of a given wordform, a vector representation reflecting both that wordform's statistical distribution in the training corpus, as well as the immediate context in which that word appears. That is, rather than producing a single, static embedding for a given string, as earlier distributional semantic measures like LSA and HAL do, BERT's contextualized embeddings are sensitive to the linguistic context in which a word token is observed. BERT's architecture appears to naturally

encode a number of linguistic features, such as part of speech, semantic roles, and others (Tenney et al, 2019). These contextualized embeddings have been shown to improve performance on a number of downstream NLP tasks involving lexical ambiguity, such as Word Sense Disambiguation (Aina et al, 2019; Loureiro et al, 2020). Past work also suggests that BERT can be used to distinguish monosemous and polysemous words, or even polysemy and homonymy (Haber & Poesio, 2020a; Haber & Poesio, 2020b; Soler & Apidianaki, 2021; Nair et al, 2020), and that BERT's representations encode sense-like information (Karidi et al, 2021). Most relevantly for our purposes, BERT's contextualized embeddings are well-suited for measuring contextual distance in a graded manner—given two contextualized embeddings of an ambiguous target word (e.g., for "marinated *lamb*" and "friendly *lamb*"), we can compute the cosine distance between those vectors, a metric often used to assess proximity in vector-space[4]. Smaller cosine distances indicate that the embeddings are closer, while larger values indicate they are further apart.

Accounting for contextual distance in a primed sensibility judgment task allows us to adjudicate among the theories outlined above. Pure Exemplar Theory predicts that the difficulty in transitioning between two contexts of use should be affected solely by their proximity in usage-space—thus, the existence of a sense boundary (or the distinction between homonymy and polysemy) should not predict variance in RT or Accuracy above and beyond cosine distance. Both varieties of the Mental Dictionary Framework predict the opposite, i.e., cosine distance should *not* explain variance in RT or Accuracy above and beyond the existence of a sense boundary. And both hybrid theories predict a systematic distortion of this continuous usage-

[4] Note that we also replicated the primary analyses using ELMo, another well-known contextualized language model. Our pre-registered analyses used BERT because it tends to outperform ELMo on Word Sense Disambiguation tasks (Wiedemann et al, 2019) and predicting relatedness judgments (Trott & Bergen, 2021), and because it was more predictive of response time in a pilot study.

space, such that the existence of a sense boundary (or the distinction between homonymy and polysemy) should increase measures relating to processing cost (e.g., RT or Accuracy), above and beyond the cosine distance as measured by BERT.

## Experiment 1

The primary goal of Experiment 1 was adjudicating between the competing theoretical accounts outlined above. As noted above (see also *Table 1*), each theory makes distinct predictions about which variables should predict processing ease (as measured by RT and Accuracy) in a primed sensibility judgment paradigm. Specifically, Pure Exemplar Theory predicts that only the continuous distance between two contexts of use is necessary to explain processing ease. In contrast, both Mental Dictionary Theories predict that only categorical variables, such as the existence of a sense boundary or the distinction between polysemy and homonymy, are necessary to explain behavior. The two hybrid theories (Hybrid and Hybrid+) subscribe to a continuous model of word meaning, but additionally hypothesize the existence of discrete sense representations (Hybrid) and a distinction between homonymy and polysemy (Hybrid+); thus, each predicts that a unique combination of these variables will predict processing ease.

This work is (to our knowledge) the first attempt to directly test the Continuity of Meaning Framework using a measure of online processing ease. Past work (Nair et al, 2020; Trott & Bergen, 2021; Li & Joanisse, 2021) has used NLM-derived measures (e.g., cosine distance) to predict relatedness judgments, but has not directly pitted those continuous measures against categorical factors (such as the existence of a sense boundary) to ask whether both explain independent sources of variance in processing difficulty.

The experimental design, hypotheses and analyses were pre-registered on OSF in advance of data collection (https://osf.io/gj48a). Additionally, data and code to reproduce the pre-registered analyses is available on OSF (https://osf.io/2s7mg/); additional data and code to reproduce the supplementary analyses is also available on GitHub (https://github.com/seantrott/trott_ph_amb).

The study was carried out with IRB approval.

**Methods.**

**Participants**. We recruited 216 participants from the UC San Diego Psychology Department Subject Pool. After following the exclusion criteria listed in our pre-registration (https://osf.io/gj48a), we had a total of 180 participants (our target sample size). The exclusion criteria included: participants who self-reported as non-native speakers of English, participants who failed at least one of the two "bot check" questions at the beginning of the experiment, participants who self-reported as having completed the experiment on a mobile device, and participants for whom more than half of critical trials were excluded because of overly slow (RT > 3 SD above the subject-level mean) or overly fast (<500 ms) responses. Of the final set of participants, 144 self-identified as female (33 male, 2 non-binary, and 1 preferred not to answer). The average age was 20.5 (SD = 1.67) and ranged from 18 to 29.

The target sample size of 180 was based on a pilot study with 74 participants. In the pilot study, we detected significant ($p < .001$) effects of both Cosine Distance and Sense Boundary, but only a marginally significant interaction between Sense Boundary and Ambiguity Type in predicting Accuracy. Thus, it was inconclusive from the pilot whether Hybrid Meaning Theory or Hybrid+ Theory was a better explanation of the data. We conducted a simulation-based power

analysis using the simR package (Green & MacLeod, 2016) to determine the number of participants we would need to detect the interaction between Ambiguity Type and Sense Boundary with 95% power at an alpha of .025 (to correct for the two dependent variables). The power analysis indicated that 95% power could be achieved with 180 participants; we then estimated the number of participants we would need based on applying the exclusion criteria to the pilot data. (More details are included in the pre-registration.)

**Materials**. We adapted materials from several previous studies (Klepousniotou, 2002; Brown, 2008; Klepousniotou et al, 2008; Klepousniotou & Baum, 2007). These studies either used sentence fragments containing an ambiguous word (e.g., "marinated *lamb*" or "*fixed* the radio"), or used homonymous and polysemous words in isolation (e.g., "bat"). For each ambiguous word, we created four sentences (two for each of the primary senses). Thus, there were six possible sentence pairs for each word: two Same Sense pairs, and four Different Sense pairs. Each sentence for each word contained the same sentence frame (e.g., "They liked the ___ lamb"), but differed in the disambiguating word (e.g., "marinated" vs. "friendly"); a minority of words (13) had at least one sentence which used a different article before the disambiguating word than the other sentences (e.g., "a" vs. "an"). We began with 115 items total (460 sentences).

We used two dictionaries (Merriam-Webster and the Oxford English Dictionary) to determine whether the two meanings expressed by a word were categorized by lexicographic experts as different senses. There were 3 words for which neither dictionary listed the meanings as separate senses at all (e.g., "glossy magazine" vs. "sports magazine"), suggesting that lexicographers viewed these meanings as the same. These items were included in the norming study, but not in the final stimulus set (leaving us with 112 words). We also used both

dictionaries to annotate whether Different Sense items were classified by lexicographers as related via Homonymy or Polysemy; meanings listed as separate entries were annotated as Homonymy, and those listed in the same entry were annotated as Polysemy. There was one word ("drill") for which the two dictionaries did not agree; in this case, we labeled the two meanings as homonymy, following the OED.

We also created a number of filler items (112 unique wordforms). Each filler word was matched for the concreteness, frequency, part of speech, and length (number of syllables) of one of the critical wordforms. Then, for each filler, we constructed two sentences containing that word, i.e., a minimal sentence pair. For 38 of these filler items (approximately one third), both sentences were nonsensical; for the remaining 74 (approximately two thirds), only one of the two sentences was nonsensical (counterbalanced for whether the first or second was nonsensical). This was to prevent participants from learning any contingencies between the prime and target item.

Finally, we ran a norming study to obtain relatedness judgments for all of the critical sentence pairs (Trott & Bergen, 2021). Eight of the words had very low relatedness judgments for their Same Sense pairs, so we excluded these from the final stimulus set, leaving us with 104 wordforms (and 624 unique sentence pairs, not accounting for order). In this final set, 30 wordforms were labeled as Homonymous, and 74 were Polysemous. 76 of the target wordforms were used as nouns and 28 were used as verbs.

Among this final set of 104 words, Mean Relatedness from the norming study was (as expected) higher among Same Sense ($M = 3.53$, $SD = 0.451$) than Different Sense ($M = 1.38$, $SD = 1.13$) pairs. Further, Different Sense Homonyms were less related on average, and also exhibited less variability ($M = 0.44$, $SD = 0.37$), than Different Sense Polysemes ($M = 1.76$, SD

= 1.11). This was also expected: the polysemous meanings ranged considerably in their relatedness, from highly related meanings (e.g., "marinated *lamb*" vs. "friendly *lamb*") to less related meanings (e.g., "brain *cell*" vs. "prison *cell*"). Additional details about the norming procedure can be found in Trott & Bergen (2021); note that some of the descriptive statistics will differ from those presented here, given that Trott & Bergen (2021) report analyses on the original set of 112 words.

**Procedure**.

Participants completed the study online. They were told that they would read a series of sentences, and that some of these sentences would make sense, while others would not. Their task was to determine which sentences made sense and which did not; they were told to indicate this via button-press (**m** for "makes sense", and **x** for "does not make sense"). The instructions encouraged participants to complete each trial as accurately and quickly as possible. Before beginning the primary experiment, participants completed ten practice trials (five sentence pairs). After each trial, they were given feedback indicating whether their response was correct.

The primary experiment contained 56 critical sentence pairs, randomly sampled from the list of possible trials. Each sentence pair contained an overlapping word (e.g., "lamb") and sentence frame ("They liked the ___"), with one disambiguating word (e.g., "marinated/friendly"). The sampling process was constrained so as not to repeat the same word multiple times across sentence pairs. A similar process was implemented for sampling 56 filler sentence pairs as well. On any given trial (i.e., a target sentence), a participant saw a sentence appear in the center of the browser page (e.g., "They liked the marinated lamb"). A reminder of their task instructions appeared below the target sentence ("Does this sentence make sense? X = No; M = Yes").

After completing the primary experiment, participants answered several demographic questions, regarding their self-identified gender, age, whether or not they were a native speaker of English, and whether or not they completed the experiment on a mobile device.

The experiment was implemented in JsPsych, version 6.0.5 (de Leeuw, 2015).

**Results**.

All analyses described below were conducted in R version 3.6.3 (R Core Team, 2020). Mixed effects models were constructed using *lmer* (for Reaction Time data) and *glmer* (for Correct Response data) commands from the *lme4* package (Bates et al, 2015). Random effects structure was determined by beginning with the maximal model, then reducing random effects as needed for model convergence (Barr et al, 2013); in this case, all models contained by-subject random slopes for the effects of Cosine Distance, Sense Boundary, and Ambiguity Type, as well as random intercepts for subjects and items. All models also contained the following covariates relating to the target word: Concreteness, Log Frequency, Part-of-speech, and Length (number of characters). Nested models were compared using log-likelihood ratio tests. Finally, each explanatory variable of interest (e.g., Cosine Distance) was used in two separate analyses, to predict either Reaction Time (RT) or Correct Response (Correct vs. Incorrect); thus, we corrected for multiple comparisons using the Holm-Bonferroni method (Holm, 1979). Only adjusted p-values are reported below.

Planned analyses were pre-registered on OSF (https://osf.io/gj48a); all exploratory analyses are marked as such in a separate section.

**Planned analyses.** First, we compared a model with fixed effects for Sense Boundary and Cosine Distance to a model omitting only the fixed effect of Sense Boundary. The full model

had significantly better fit than the reduced model for both Accuracy [$X^2(1) = 77.17, p < .001$] and RT [$X^2(1) = 34.43, p < .001$]. This disconfirms the prediction of Pure Exemplar Theory; even after adjusting for continuous differences in a word's context of use, the existence of a sense boundary explained additional variance in how accurately and quickly participants responded to the target item. Subjects were more likely to respond correctly to Same Sense items (89.3%) than Different Sense items (79.9%). Response times were also faster for Same Sense ($M = 1068, SD = 542$) than Different Sense ($M = 1159, SD = 598$) items.

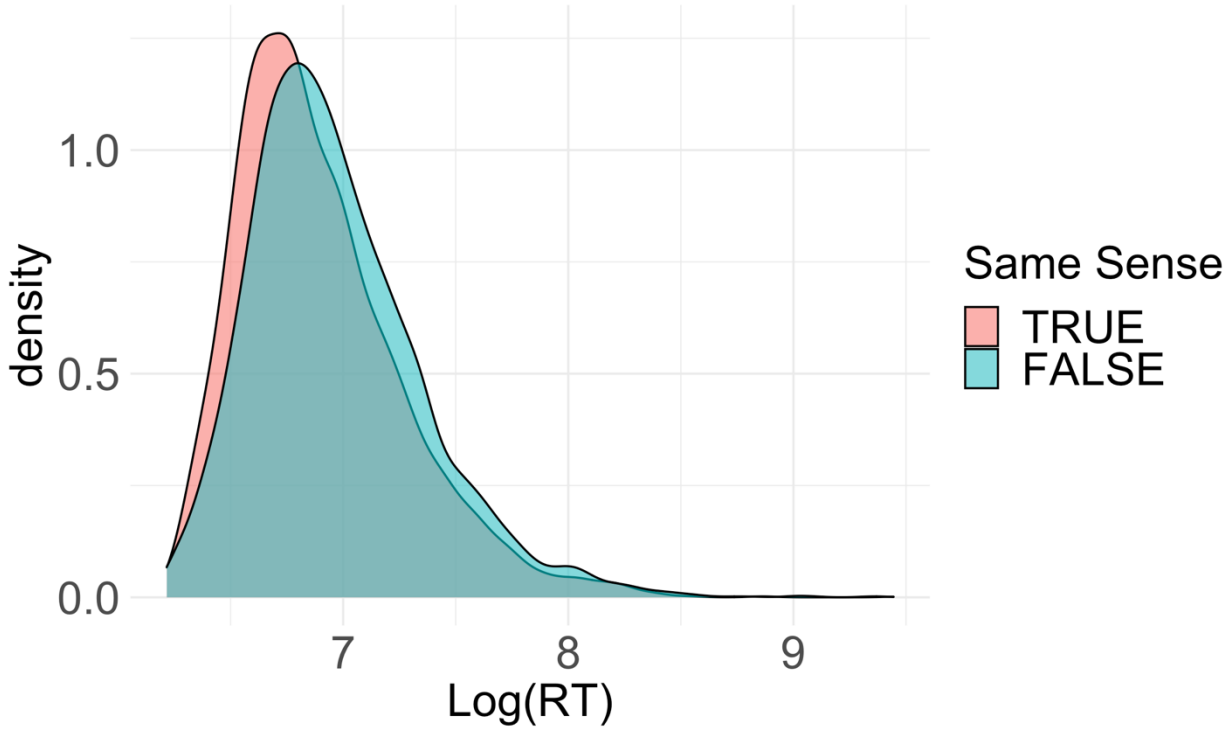


Figure 12: Log Reaction Time for correct trials only, displayed as a function of Same Sense vs. Different Sense. Different Sense trials resulted in longer response times on average than Same Sense trials.

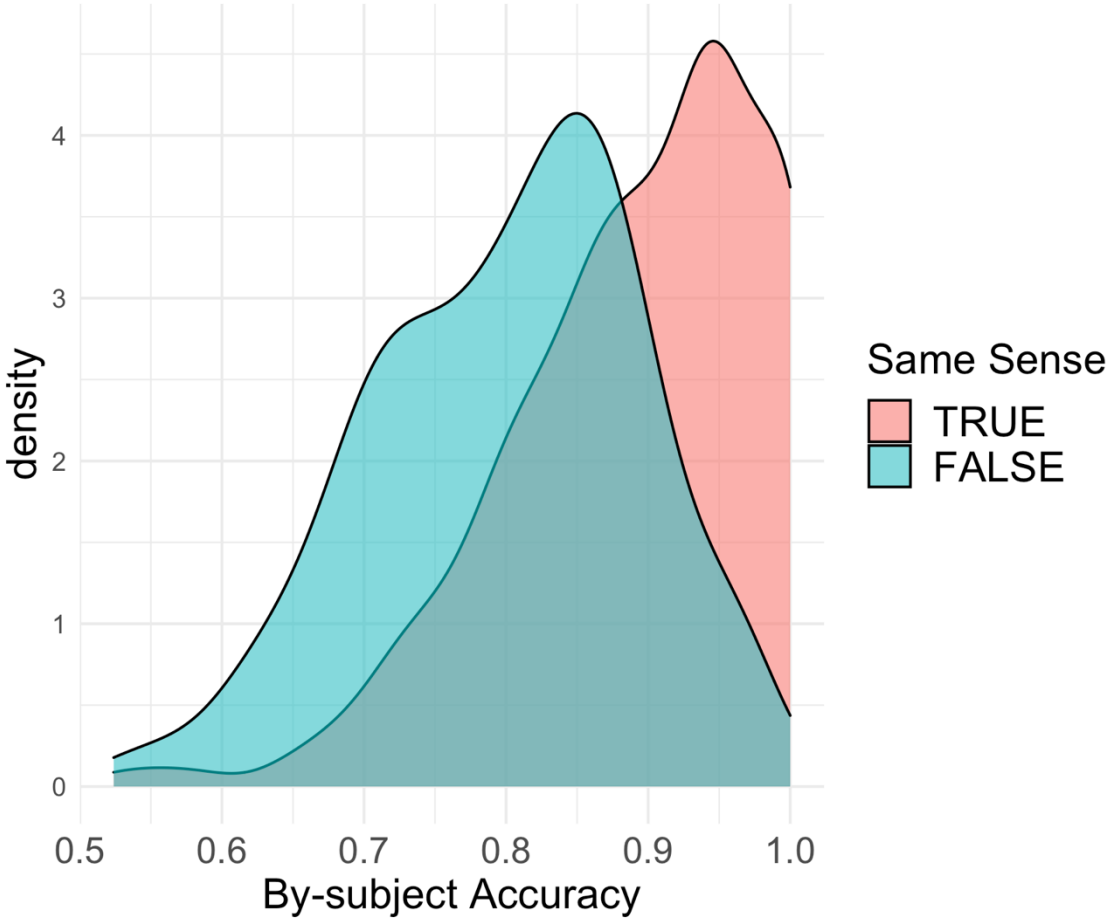


Figure 13: Accuracy on the target trial, grouped by subject and displayed by Same Sense vs. Different Sense. Accuracy was considerably higher for Same Sense (M = 0.89) than Different Sense trials (M = 0.8).

Second, we constructed a full model including fixed effects of Sense Boundary, Cosine Distance, and Ambiguity Type, as well as an interaction between Ambiguity Type and Sense Boundary. The full model explained significantly more variance in RT than a model omitting only the fixed effect of Cosine Distance [$X^2(1) = 15.42, p < .001$]; larger Cosine Distances were associated with longer response times [$\beta = 0.14, SE = 0.04$]. This disconfirms predictions of both theories falling under the Mental Dictionary Framework, i.e., both the Core Representation and Sense Enumeration accounts; continuous gradations in a word's context of use predicted

behavior above and beyond Sense Boundary and Ambiguity Type. There was no significant effect of Cosine Distance on Accuracy ($p > .2$).

Finally, we compared the full model to a model omitting only the interaction between Ambiguity Type and Sense Boundary. The full model did not explain significantly more variance for either Accuracy ($p = .16$) or RT ($p > .2$). This fails to disconfirm Hybrid Meaning Theory—at least on this task, there is no evidence that homonymy and polysemy elicit fundamentally different behavior, all other things being equal.

**Exploratory analyses.** One unexpected finding was an apparent main effect of Ambiguity Type on Accuracy—participants were considerably less accurate when responding to homonymous than polysemous items, as evidenced by a significant coefficient for Ambiguity Type [$\beta = -1.14, SE = 0.24, p < .001$]. Because this effect occurred in both the Same Sense and Different Sense conditions, it is unlikely to be driven by relative differences in the degree of cross-sense inhibition (or facilitation) across Polysemy and Homonymy. That is, the main effect cannot be due to priming. Further support for this interpretation comes from inspection of Accuracy on the first, unprimed half of each trial, which reveals a similar main effect of Ambiguity Type (but not, crucially, of Sense Boundary). Together, this suggests that the main effect of Ambiguity Type arises because of properties of the sentences themselves—either because participants are less accurate when responding to sentences with homonyms in general, or because these *particular* sentences were not sufficiently controlled for plausibility across Homonymy and Polysemy.

To account for the latter possibility, we replicated the planned analyses from above, but substituted mean first-trial Accuracy (or RT) for each version of each item in place of variables relating to the target word like Concreteness or Frequency. By including first-trial Accuracy (or

RT) as a covariate, we could better estimate parameters of interest (e.g., coefficients for Cosine Distance, Sense Boundary, and the Ambiguity Type x Sense Boundary interaction) as they relate to the task itself, as opposed to uncontrolled properties of the stimuli. The main effect of Ambiguity Type in the full model was now not significantly different from zero ($p > .5$ for both Accuracy and RT). Importantly, however, the main effect of Sense Boundary was preserved for both Accuracy [$X^2(1) = 84.8, p < .001$] and RT [$X^2(1) = 36.57, p < .001$], as was the main effect of Cosine Distance on RT [$X^2(1) = 17.73, p < .001$]. Again, there was no significant interaction between Ambiguity Type and Sense Boundary for either RT or Accuracy ($ps > .2$).

While this does not answer directly the question of why Homonymous sentences had lower accuracy rates than Polysemous sentences overall, it does suggest a method for directly accounting for any uncontrolled differences in the stimuli.[16] This approach has the advantage of more directly adjusting for any variance due to features intrinsic of the individual sentence in question; differences in First-Trial Accuracy or First-Trial RT are not plausibly attributed to the structure of the task—given that the target ambiguous word has not been directly primed or inhibited by a previous use—and instead, reflect processing difficulties relating to the sentence itself. Thus, in Experiment 2, we sought to replicate the findings reported in Experiment 1 using this refined analysis.

## Experiment 2

In Experiment 1, we found that behavior was predicted both by Cosine Distance and Sense Boundary, but not by the interaction between Sense Boundary and Ambiguity Type. However, there was a main effect of Ambiguity Type: accuracy was lower for homonymous than polysemous sentences. This main effect could have arisen from uncontrolled properties of the

[16] See the General Discussion for possible explanations for this result.

stimuli—indeed, when we controlled for the first-trial accuracy of each item, the main effect of Ambiguity Type disappeared, but the main effects of Sense Boundary and Cosine Distance were preserved. However, this analysis was exploratory. Thus, the primary goal of Experiment 2 was replicating the main findings of Experiment 1 while pre-registering this new analysis (https://osf.io/4ej6t).

**Methods**

**Participants**. As in Experiment 1, we aimed to collect data from 180 participants. Rather than try to estimate the rate of exclusion ahead of time, we iteratively collected data in batches and applied the exclusion criteria to each batch until *at least 180* included participants were reached.

Subjects were recruited through the UC San Diego Psychology Department Subject Pool. When we finished collecting data, there were 239 subjects in the final pool, with 187 remaining after applying the exclusion criteria. Of the final 187 participants, 129 self-reported as female (53 male, 2 non-binary, and 3 preferred not to answer). The average age was 20.4 (SD = 2.04), and ranged from 18 to 32.

**Materials and Procedure**. The materials used and experimental design were identical to Experiment 1.

**Results**

**Planned analyses.** The analyses were identical to those carried out in Experiment 1, except that the lexical statistics of the target word (e.g., Concreteness or Log Frequency) were replaced by the average first-trial accuracy (or RT) for the target sentence.

As in Experiment 1, predictions of Pure Exemplar Theory were disconfirmed by finding a significant effect of Sense Boundary above and beyond Cosine Distance, for both Accuracy

[$X^2(1) = 96.96, p < .001$] and RT [$X^2(1) = 45.57, p < .001$]. Predictions of both theories falling under the Mental Dictionary Framework were also disconfirmed by finding a significant effect of Cosine Distance above and beyond Sense Boundary and Ambiguity Type (and their interaction), when predicting RT [$X^2(1) = 39.64, p < .001$] but not Accuracy ($p > .2$). Finally, we detected no significant interaction between Ambiguity Type and Sense Boundary for either Accuracy or RT ($p$s $> .2$).

**Exploratory Analyses**. The analyses above, precisely like the results from Experiment 1, are most consistent with Hybrid Meaning Theory. However, as noted in the Introduction, there are multiple mechanisms by which sense categories could be implemented in a continuous space. In the Sense Attraction Account, distances in usage-space are *reduced* for within-sense tokens, and *exaggerated* for tokens that span a sense boundary. Crucially, within-cluster variance is not eliminated entirely—it is merely reduced. In contrast, the Sense Distillation Account claims that within-cluster variance is entirely distilled into a single point, i.e., the centroid of that cluster. The metric properties of the underlying continuous space are preserved across sense-clusters, but within-cluster variance is removed.

These accounts make testable predictions about whether, and how, Cosine Distance is related to processing ease for Same Sense items. Specifically, the Sense Attraction Account predicts that even for Same Sense items, reaction time should increase as a function of Cosine Distance (as it does when all items are considered). However, because the Sense Distillation Account claims that within-cluster variance is removed entirely, it predicts that Cosine Distance should not be systematically related to reaction time.

We tested these accounts by building a linear mixed effects model with Log RT as a dependent variable, fixed effects of both Cosine Distance and Ambiguity Type, by-subject

random slopes for both Cosine Distance and Ambiguity Type, and random intercepts for subjects and items. This model explained significantly more variance than a model omitting Cosine Distance alone [$X^2(1) = 11.02$, $p = .001$], indicating that even within Same Sense pairs, Cosine Distance was positively correlated with RT [$\beta = .17$, $SE = 0.05$]. This disconfirms the Sense Distillation Account, which predicts no difference in RT within Same Sense pairs.

**Discussion**

Combined with Experiment 1, these results are inconsistent with three of the five accounts under investigation. As noted earlier, each account made specific predictions about which variables should or should not influence behavior in a primed sensibility judgment task (see *Table 1* for a summary). Only the Hybrid+ Theory could not be strictly disconfirmed, given that it predicts significant effects of all the relevant experimental variables; failing to find a significant effect is not necessarily grounds for rejecting a theory. Nevertheless, a simpler theory that explains the data equally well is still preferable from the standpoint of theoretical parsimony. In this case, that reasoning tips the scales toward the Hybrid Meaning Theory.

To summarize the results, first, we found that the existence of a sense boundary between two contexts of use (e.g., "marinated *lamb*" vs. "friendly *lamb*") resulted in slower response times and less accurate responses overall, as compared to two contexts of use that fall under the same sense category (e.g., "marinated *lamb*" vs. "roast *lamb*"). This replicates the sense consistency effect obtained in past work, using both identical task paradigms (Klein & Murphy, 2001; Yurchenko et al, 2020) and alternative approaches (Klein & Murphy, 2002). Importantly, this effect held even after controlling for contextual distance, disconfirming the prediction of Pure Exemplar Theory. That is, behavior on this task can be better explained by positing some

form of categorical sense representation above and beyond the distance between two contexts of use.

Second, we found that response times were systematically longer for larger contextual distances, as measured by the cosine distance between BERT's contextualized representations of the ambiguous target word, when controlling for sense boundaries. This disconfirms predictions of both accounts falling under the Mental Dictionary Framework (i.e., the Core Representation and Sense Enumeration Accounts), neither of which allow for graded effects of context: behavior on this task varied not only as a function of discrete sense representations, but rather was related to a measure that captures the context-dependent nature of word meaning. To our knowledge, this is the first empirical demonstration that online processing difficulty of ambiguous words can be explained by a continuous measure of contextual distance, above and beyond discrete variables like Sense Boundary.

This leaves the two hybrid theories: Hybrid vs. Hybrid+. The former predicts no difference in behavior across polysemous and homonymous words, while the latter does. Crucially, we failed to detect a difference in how people processed different-sense polysemous meanings and different-sense homonymous meanings, after controlling for differences in first-trial accuracy or response time. Although we cannot strictly reject the Hybrid+ Theory—absence of evidence does not entail evidence of absence—this does suggest that the Hybrid Meaning Theory is a more parsimonious explanation of the data from both experiments.[17] According to this theory, the clusters in context-space that arise as a function of purely distributional properties of language use are systematically "warped" in psychological space, as in the categorical perception of speech (Goldstone & Hendrickson, 2010), according to sense

[17] The question of whether homonymy is truly just a form of "distant" polysemy is further explored in the General Discussion.

boundaries. Further, a post-hoc analysis of the data from both experiments found that variance in contextual distance within Same Sense words (i.e., within a sense category) was also predictive of reaction time. This suggests that of the two compression mechanisms explored in the Introduction (Sense Attraction vs. Sense Distillation), Sense Attraction is a better explanation of the data.

This result raises a number of questions about the nature of these sense representations. How exactly does *contextual distance* map onto *conceptual distance*? Which functional transformation best accounts for the behavioral data, and what are the parameters underlying this transformation? These questions are explored in the section below.

**Hybrid Meaning Theory: A Further Test and Computational Model**

Above, we concluded that Hybrid Meaning Theory—and the Sense Attraction mechanism in particular—was the best explanation of the behavioral data. This theory claims that distance in *context-space* is systematically warped by the existence of sense boundaries, such that within-sense distances are reduced, and across-sense distances are amplified.

One potential objection to this conclusion is that BERT's representation of the context space is not analogous to that of human participants. In principle, it is possible that human meaning representations are completely continuous (as predicted by Pure Exemplar Theory)—and even derived from distributional statistics alone—but that the topology of this representational space is distinct from BERT's, for reasons other than the existence of putative sense boundaries. If this interpretation is correct, BERT's representational space already contains sufficient information to account for human behavior, provided it is transformed in the appropriate way. Critically, to be consistent with Pure Exemplar Theory, such a transformation must be *bottom-up*: that is, it must not depend on information extrinsic to what is observable via

a word's pattern of use. On the other hand, if Hybrid Meaning Theory is correct, no bottom-up transformation to the underlying BERT-space will be sufficient to account for human sense knowledge. Instead, contextual distance must be transformed using *top-down* or auxiliary information—i.e., information that cannot be derived from distributional regularities in linguistic input alone. For example, contextual distance could be systematically transformed according whether the two contexts of use straddle a sense boundary or not.

In the current section, we asked whether a top-down or bottom-up transformation to cosine distance improved the fit of a model predicting human behavior on the primed sensibility judgment task. Specifically, we compared the success of several bottom-up transformations to top-down transformations relying on the value of the Sense Boundary parameter (i.e., Same Sense vs. Different Sense). As a second-order question, we also considered two distinct functions to apply to cosine distance (for both bottom-up and top-down transformations): 1) an additive function, which increased or decreased cosine distance as a function of Sense Boundary (or the induced cutoff parameter); and 2) a multiplicative function, which scaled with the original value of cosine distance. Both functions, as well as the procedure for identifying the optimal parameters for each transformation, are described in more detail in the Methods section below.

Once the parameters for each transformation were identified, we asked which transformation best predicted human behavior on Experiments 1-2. The best transformation was then selected using Akaike Information Criterion (AIC), a measure of model fit (Akaike, 1974; Burnham & Anderson, 2002). That is, we compared the predictive power of a series of statistical models, each containing a specific implementation of Transformed Distance.

**Functional Transformations**

The first functional transformation was additive and top-down. That is, it assumed a fixed mapping between contextual distance and conceptual distance, according to whether or not two contexts of use were separated by a lexicographer-classified sense boundary. This mapping can be described as follows:

$$Y = \begin{Bmatrix} same = 1 : x - \beta_1 \\ same = 0 : x + \beta_1 \end{Bmatrix}$$

If two contexts of use correspond to the same sense, this function decreases conceptual distance by a fixed amount[18] ($\beta_1$); if two contexts of use correspond to distinct senses, this function increases conceptual distance by a fixed amount ($\beta_1$). The "bottom-up" version of this transformation is identical, but uses an optimized cutoff parameter instead of Sense Boundary:

$$Y = \begin{Bmatrix} x \leq c : x - \beta_1 \\ x > c : x + \beta_1 \end{Bmatrix}$$

The second functional transformation was still linear, but no longer applied a fixed transformation to a given value of Cosine Distance. Rather, Transformed Distance was scaled proportionally to the original value of Cosine Distance: for same sense pairs, more distant pairs were "attracted" more relative to closer pairs; for different sense pairs, closer pairs were "repelled" more relative to already distant pairs. This was based on research suggesting that certain category effects are particularly large near category boundaries (Kuhl, 1991), and that co-categorized exemplars undergo a larger perceptual transformation when they are further apart (Kuhl, 1991). This mapping can be described as follows:

[18] Using a single same term ($\beta_1$) produces the same optimal solution as using distinct terms ($\beta_1$, $\beta_2$) for same and different sense pairs.

$$Y = \begin{Bmatrix} same = 1: \frac{x}{\beta_1} \\ same = 0: x + \frac{1 - x}{\beta_2} \end{Bmatrix}$$

That is, the contextual distance of same sense pairs is divided by a fixed amount ($\beta_1$), which results in a proportionately larger transformation to distant pairs than close pairs. Conversely, the contextual distance of different sense pairs is increased by an amount that decreases as Cosine Distance increases—different sense pairs that are already very distant (i.e., close to 1) will be adjusted less than different sense pairs that are very close (i.e., close to 0). As with (1), the bottom-up version of this transformation uses an optimized cutoff parameter instead of the Sense Boundary variable:

$$Y = \begin{Bmatrix} x \leq c: \frac{x}{\beta_1} \\ x > c: x + \frac{1 - x}{\beta_2} \end{Bmatrix}$$

For each transformation, we performed a grid search over a constrained parameter space to identify the optimal set of parameters that would best approximate relatedness. For the additive transformation, we considered values of $\beta_1$ ranging from a lower-bound of 0 (i.e., no transformation) to an upper-bound of 1. For the multiplicative transformation, we considered parameter values ranging from [.1, 15] for both $\beta_1$ and $\beta_2$. For the bottom-up versions of each transformation, we considered cutoff parameters between [0, 1].

**Parameter Optimization**

To determine the optimal values of each parameter for each functional transformation, we sought to optimize the strength of the relationship between Transformed Distance and human relatedness judgments. Past work (Trott & Bergen, 2021) has found that although Cosine Distance is strongly correlated with relatedness (*rho = -.58*), it underperforms human inter-annotator agreement by a considerable margin (*rho = -0.79*); further, Cosine Distance systematically *underestimates* human relatedness judgments of same-sense pairs, and *overestimates* the relatedness of different-sense pairs. Thus, we used a grid search to identify the parameters for each transformation that optimized the correlation strength between Transformed Distance and Mean Relatedness.

The optimal parameters and resulting correlations between Mean Relatedness and Transformed Distance are included in *Table 2*, and the transformations themselves are depicted in the figures below.

Table 2: Final parameter values for each transformation.

| **Transformation** | **Parameters** | **Pearson's *r*** |
|---|---|---|
| Additive (BU) | $\beta_1 = 0.2$, $C = 0.2$ | -0.59 |
| Multiplicative (BU) | $\beta_1 = 0.6$, $\beta_2 = 14.6$, $C = 0.5$ | -0.62 |
| Additive (TD) | $\beta_1 = 0.4$ | -0.76 |
| Multiplicative (TD) | $\beta_1 = 10.6$, $\beta_1 = 3.6$ | -0.77 |

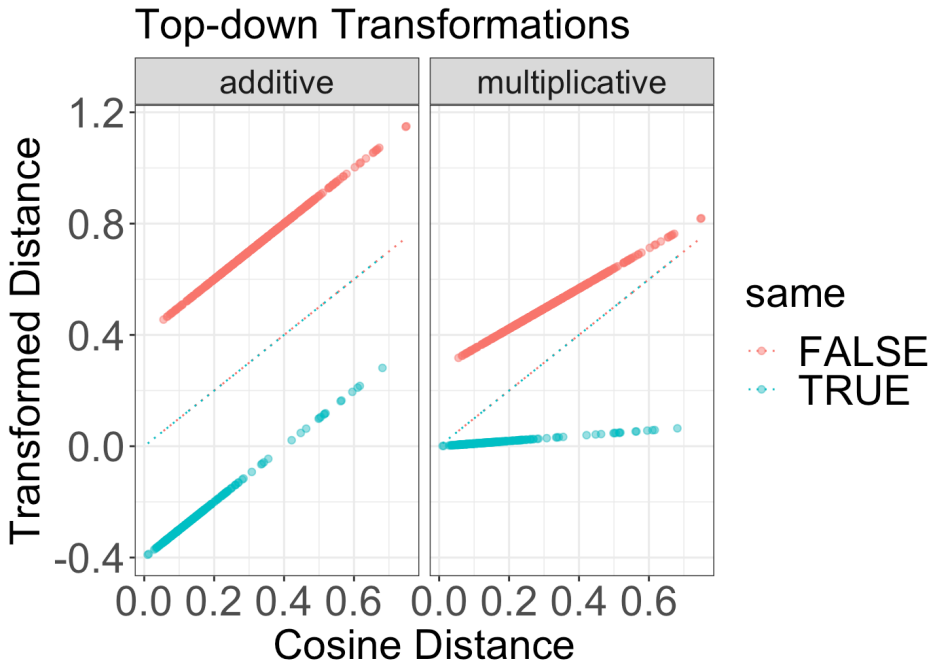


Figure 14: Final result of top-down transformations to Cosine Distance. Different functional transformations are applied to Cosine Distance as a function of Sense Boundary.

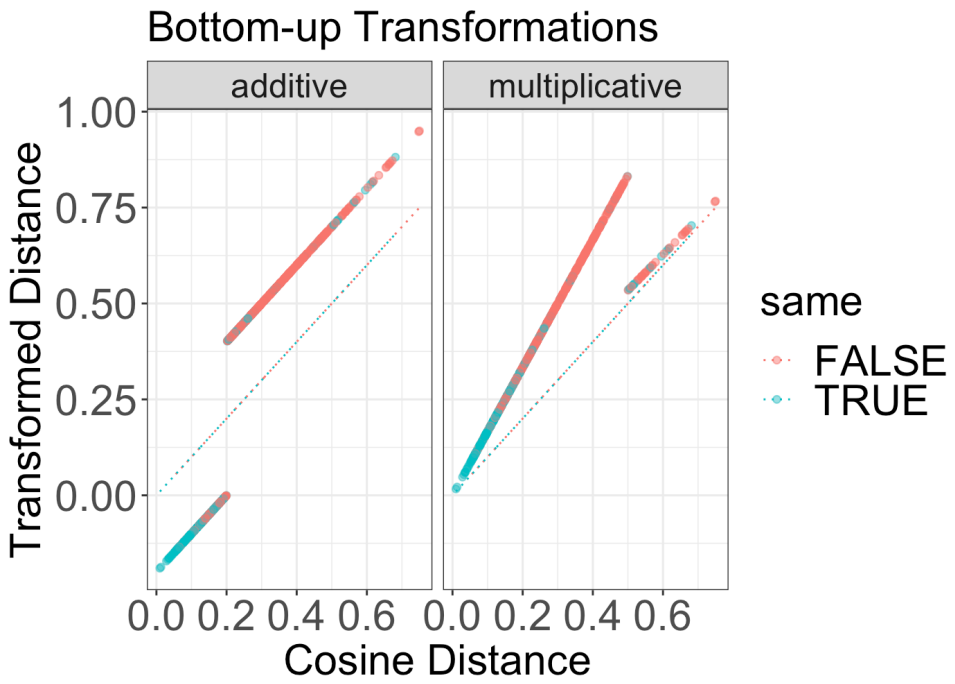


Figure 15: Final result of bottom-up transformations to Cosine Distance. Here, distinct transformations are applied according to some cutoff parameter, as opposed to the value of Sense Boundary variable; the cut-off refers to the value of Cosine Distance at which to apply one transformation or the other, and stands in for a "bottom-up" or induced value of Sense Boundary. Note that in the case of the multiplicative transformation, this results in a surprising transformation: because some Different Sense pairs are grouped under the cutoff value, there is a pressure to increase the distance of those pairs even more, to differentiate them from the Same Sense pairs also grouped under the cutoff value.

**Model Specification and Evaluation**

Our primary goal was to identify the transformation that best predicted human behavior on Experiments 1-2. To this end, we compared a series of models with distinct parameterizations, predicting both Correct Response and RT. Recall that in both Experiments 1-2, Cosine Distance did not improve model fit when predicting Correct Response, but it did explain independent

variance in RT; Sense Boundary explained variance in both dependent variables. There were four transformed models total, accounting for the transformation itself (Additive vs. Multiplicative) and the implementation (Bottom-Up vs. Top-Down). All models included the same random effects structure, and differed only in which fixed effects were added.

1. Transformed Distance (Additive):
    *a.* D-Add-BU
    *b.* D-Add-TD
2. Transformed Distance (Multiplicative)
    *a.* D-Mul-BU
    *b.* D-Mul-TD
3. Original Cosine Distance: D
4. Sense Boundary: SB
5. A model containing both (3) and (4): D + SB
6. A model containing an interaction between D and SB, along with their main effects.

The top-down additive and multiplicative models (D-Add-TD and D-Mul-TD) represent hypothesized implementations of the Sense Attraction Account, i.e., distinct mechanisms by which within-sense distance is reduced and across-sense distance is increased. In this sense, they are each examples of a "hybrid" model. Thus, to the extent that Cosine Distance and Sense Boundary each explain unique variance in behavior, as they do for Reaction Time, these hybrid models should improve upon models with only Cosine Distance (D) or Sense Boundary (SB). Their bottom-up counterparts are included to test whether equivalent transformations to Cosine Distance *without* the use of extrinsic information (i.e., a sense boundary) would suffice.

Model D + SB is another example of a "hybrid" model, which is agnostic to the particular transformation applied to Cosine Distance, but which simply accounts for both Sense Boundary and contextual distance using distinct parameters. If D + SB is superior to both of the models with transformed distance, it suggests that neither functional transformation is sufficient to capture the underlying psychological transformation. But if either D-Add or D-Mul improves upon D + SB, it suggests that the corresponding functional transformation is, in fact, a good approximation of the true mapping between contextual distance and conceptual distance. Finally, we considered a model with an interaction between Cosine Distance and Sense Boundary (D * SB). This model can be seen as a superset of the multiplicative transformations, since it allows for a different slope of the effect of Cosine Distance for Same Sense vs. Different Sense pairs.

Further, because there are both top-down and bottom-up implementations of D-Add and D-Mul, we can ask whether—and to what degree—an explicit, supervised transformation improves upon one that simply warps cosine distance according to some cutoff parameter. If the top-down transformations do not represent an improvement, it suggests that the relevant information to form human-like sense boundaries is already captured by the distributional regularities of language use—that is, the transformation does not require information *external* to contextual distance (as measured by BERT). Importantly, this outcome would be consistent with Pure Exemplar Theory: human lexical knowledge can be explained using information present in the distributional statistics of linguistic input alone. But if the top-down transformations do improve upon the bottom-up ones, it suggests that other sources of information, or other manners of representation, are necessary to account for human behavior.

We then calculated the AIC for each model. AIC is a measure of model fit, and is defined as:

$$AIC = 2k - 2\ln(L)$$

Where $k$ is the number of parameters in the model, and $L$ is the likelihood of the model. Models with better fit will have higher values of $L$, and thus lower AIC values overall. As is standard practice (Burnham & Anderson, 2002; Burnham et al, 2011), we rescaled each value of AIC by subtracting the AIC of the best model (i.e., the one with the lowest AIC) of that model set.

**Results**

**Predicting Response Time.** First, we considered the distribution of AIC values across models for predicting RT, aggregated across Experiments 1-2.

The best model (i.e., the one with the lowest AIC) is the model containing both of the original predictors (Cosine Distance and Sense Boundary), followed by the model containing their interaction; presumably, the interaction does not substantially improve model fit, and is penalized for adding an extra parameter.

None of the transformations considered were sufficient to account for the information provided by Cosine Distance and Sense Boundary. On the other hand, the top-down implementations of the additive and multiplicative transformations represented a substantive improvement over Cosine Distance alone, as well as Sense Boundary. This is not entirely surprising, given that the transformed variables explicitly incorporate both Cosine Distance and Sense Boundary, systematically adjusting the former as a function of the latter. Of the two transformations, the simpler Additive transformation resulted in a lower AIC than the Multiplicative transformation.

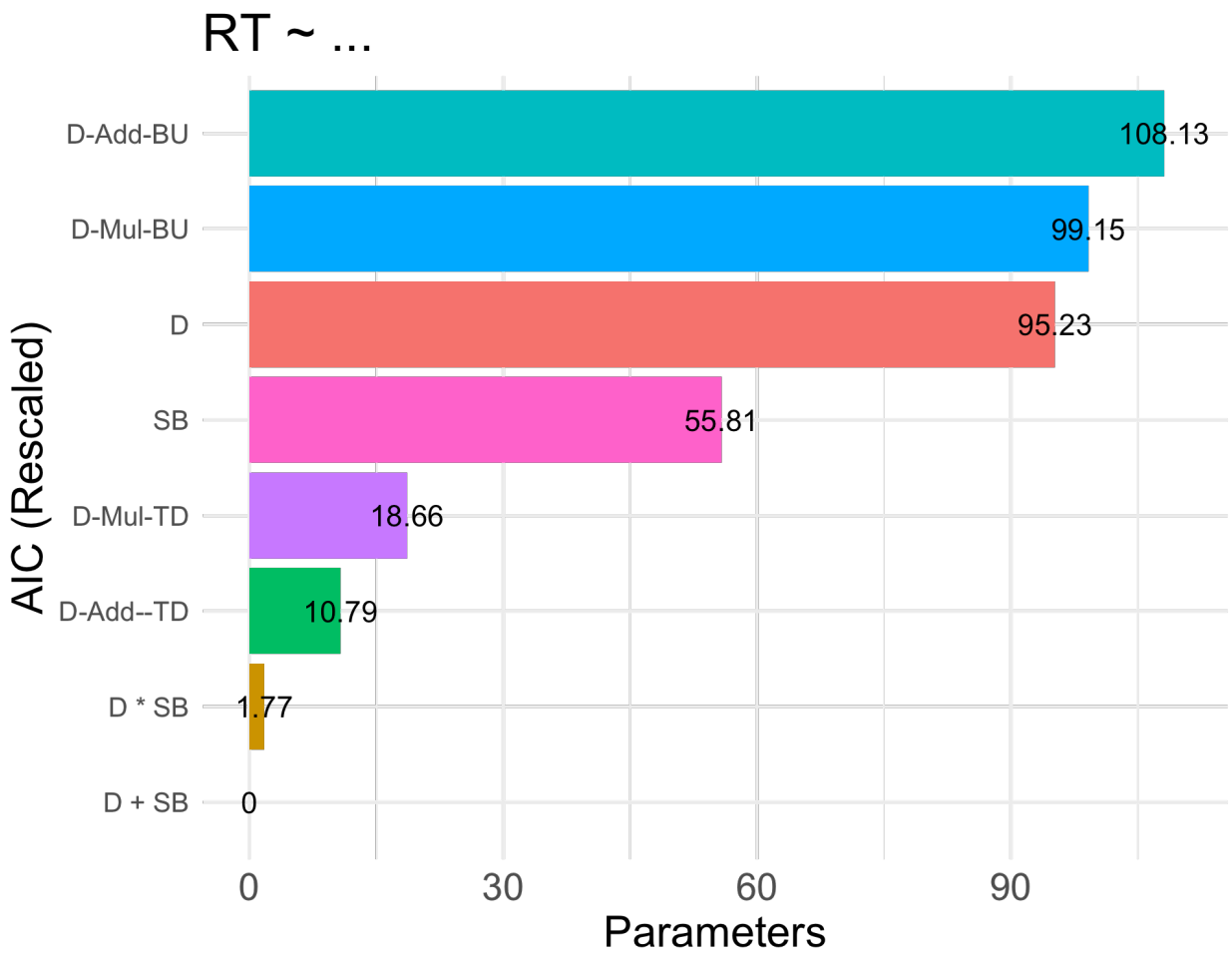


Figure 16: Rescaled AIC for each of the models predicting RT. The models containing top-down transformations (D-Add-TD and D-Mul-TD) exhibited better fit than those containing only Sense Boundary (SB) or the original Cosine Distance variable (D). The bottom-upt ransformations (D-Add-BU and D-Mul-BU) exhibited the worst fit.

Finally, the bottom-up implementations of both transformations actually performed *worse* than Cosine Distance alone. This is more surprising, given that they were optimized to improve the correlation between Cosine Distance and Mean Relatedness.

**Predicting Accuracy.** Second, we asked how well the transformed versions of Cosine Distance predicted Correct Response. Recall that in Experiments 1-2, a fixed effect of Cosine Distance did not improve model fit above a model containing only Sense Boundary.

In this case, the original measure of Cosine Distance performed the worst, followed by the bottom-up (BU) transformations; unlike with RT, the bottom-up transformations did represent an improvement upon the original Cosine Distance measure.

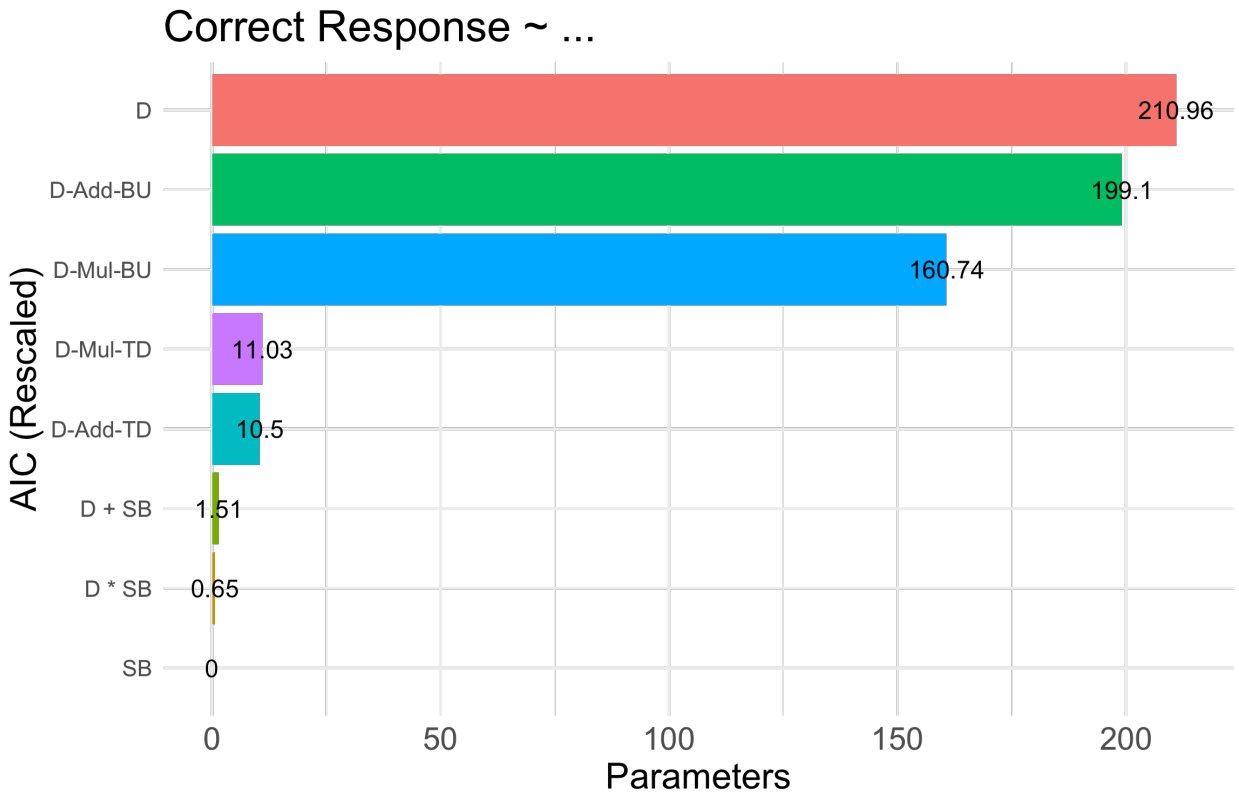


Figure 17: Rescaled AIC of the models predicting Correct Response. As with RT, the models with the top-down transformations (D-Mul-TD and D-Add-TD) exhibited better fit than those with the bottom-up transformation (D-Mul-BU and D-Add-BU), though in this case, these were not as successful as a model with Sense Boundary alone.

The best model was the one containing only Sense Boundary. Again, this is not surprising, given that the addition of Cosine Distance did not improve model fit for Experiments

1-2. Of the top-down (TD) transformations, the Additive transformation resulted in a slightly lower AIC, but this difference was quite small (~0.53), considering the differences between other models.

**Discussion**

In this section, we attempted to formalize and compare different implementations of Hybrid Meaning Theory. This theory claims that distance in context-space is systematically warped in conceptual space by the existence of sense boundaries, such that within-sense distance is reduced and across-sense distance is increased.

We considered two high-level questions. First, what information is required to account for the effect of sense boundaries? Can these effects be simulated by applying a bottom-up transformation to Cosine Distance, or does a successful approximation require some top-down, external source of information? And second, which functional transformation (i.e., additive vs. multiplicative) results in a parameter that best predicts human behavior?

We addressed the first question by comparing a top-down and bottom-up version of each transformation. The key difference was that the top-down transformations explicitly relied on the value of Sense Boundary (i.e., Same vs. Different Sense), while the bottom-up transformations induced an optimal "cutoff" parameter to apply to Cosine Distance. Models equipped with the top-down transformations consistently outperformed those using the bottom-up transformations, as measured by a lower AIC value. This pattern held across both dependent variables (Correct Response and Response Time) and both types of transformation (additive vs. multiplicative). This suggests that distributional statistics alone, at least as operationalized by certain state-of the-art Neural Language Models, are insufficient to account for the effect of sense categories.

Rather, an explanatory theory must posit that sense category structure is derived from some source of information or representation that goes beyond linguistic co-occurrence statistics; plausible candidates are explored in the General Discussion.

We addressed the second question by comparing two functional transformations. The additive transformation was intended to model a main effect of Sense Boundary: within-sense distance was reduced by some fixed amount, and across-sense distance was increased by that same amount. In contrast, the multiplicative transformation allowed the magnitude of a given transformation to vary with the original distance in context-space: distant Same Sense pairs were attracted more than pairs that were already close together; and nearby Different Sense pairs were repelled more than pairs that were already distant. When predicting RT, the top-down additive transformation was better than the top-down multiplicative transformation; this was also true when predicting Correct Response, but the difference in predictive power was comparatively very small.

## General Discussion

We began with the question of how humans store and represent the meanings of ambiguous words. Traditional theories fall under the Mental Dictionary Framework, with discrete entries corresponding to each meaning of a wordforms. In contrast, the Continuity of Meaning Framework views word meaning as trajectories in a continuous, context-dependent state-space (Elman, 2004; Li & Joanisse, 2001). Some theories falling under this framework (e.g., Pure Exemplar Theory) eschew the notion of discrete meaning representations altogether. In this paper, we also introduced two "hybrid" theories, which allow for the possibility of graded,

context-sensitive meaning representations, but also posit the existence of mediating categorical representations (see *Table 1* for a summary).

Two behavioral experiments provided support for the simpler of these two hybrid theories, which we uncreatively call Hybrid Meaning Theory. Using a primed sensibility judgment paradigm, we found that response time on the target trial was systematically related to the continuous distance between the prime and target contexts, above and beyond the existence of a sense boundary between those contexts—this disconfirms any theory that fails to account for continuous, context-sensitive meaning representations (e.g., any theory under the Mental Dictionary Framework). Both response time and accuracy were further modulated by the existence of a sense boundary, disconfirming a theory that posits no discrete meaning representations (e.g., Pure Exemplar Theory). We also found no evidence that the size of this effect depended on the kind of ambiguity (i.e., homonymy vs. polysemy), suggesting that these phenomena do not elicit categorically distinct behavior on the task. Altogether, this suggests that Hybrid Meaning Theory accounts best for the behavioral signatures we measured of how humans represent the meaning of ambiguous words.

Below, we discuss limitations of the current work, and explore implications for future research.

**Accuracy vs. Reaction Time (RT)**

As described above, both dependent measures (Accuracy and RT) were predicted by the existence of a sense boundary, but only RT was significantly correlated with cosine distance. Although we treated these measures as testing the same hypothesis (hence correcting for multiple comparisons), it is worth exploring potential post-hoc explanations for why they would diverge with respect to cosine distance.

Accuracy is a discrete measure (correct or incorrect), reflecting discrete responses on the task (sensible vs. nonsensical). Consequently, it reflects the outcome of imposing a decision threshold on a process that may, at root, be continuous. In cases where the effect of graded context distances (here, cosine distance) is relatively small—they may be detectable only on the process of *arriving* at a decision, but not necessarily the outcome of a decision itself.

In contrast, RT reflects the amount of time required to correctly identify a sentence as plausible. As a more fine-grained measure of the process by which a participant *arrived* at their decision, RT may thus be more suitable for identifying small, continuous effects like that of cosine distance. Indeed, other researchers (Spivey & Dale, 2004; Spivey & Dale, 2006) have pointed out that fine-grained, continuous measures are important for investigating putatively continuous processes. If this is true, it suggests a potential avenue for future work: researchers might deploy more fine-grained measures (e.g., mouse-tracking, eye-tracking, or EEG) to identify whether and to what extent the mental lexicon exhibits continuity.

**Limitations of the Language Model**

One possible objection to the current work is that BERT represents a poor operationalization of Pure Exemplar Theory, and that other language models would be a better choice. This objection might manifest in two different ways: first, that BERT already has representational abstractions, and is thus ill-suited to operationalizing an account that eschews sense representations (i.e., it's already too humanlike); and second, that BERT is too limited, either in its training data or its architecture (i.e., it's not powerful enough).

First, as others have pointed out (Mahowald et al, 2020), signals elicited from neural language models—e.g., surprisal, hidden unit activation, etc.—often covary with psychological or linguistic categories, such as parts of speech, animacy, semantic roles, and more (Tenney et al,

2019); this is sometimes interpreted as reflecting the formation of representational abstractions. If models like BERT are capable of forming abstractions, it is conceivable that sense representations might already be encoded by BERT, in which case BERT would indeed be a poor implementation of a theory that posits no sense representations. However, this objection can be rejected on empirical grounds: in our studies, BERT demonstrably failed to capture variance in human behavior that *was* explained by the existence of a sense boundary. Similarly, bottom-up transformations to cosine distance alone failed to improve model fit above and beyond the effect of sense boundary—the best transformations were "top-down", in that they relied on an external source of information (in this case, human-annotated sense knowledge). Together, these findings empirically demonstrate that even if BERT is capable of forming representational abstractions, these abstractions cannot account for the effect of human sense knowledge.

A second, alternative objection is that BERT is not sufficiently powerful. Neural language models are evolving rapidly—models like GPT-3 already surpass BERT on a number of metrics (Brown et al, 2020), and increases in computing resources will likely yield even better models in years to come (Kaplan et al, 2020). Thus, our "best guess" for how much information can be extracted from linguistic context alone may change as well; it is possible that a future generation of neural language models *will* display something equivalent to human sense knowledge. Importantly, however, improvements in performance along some dimensions (e.g., perplexity) do not always entail better predictions of human behavior on other tasks (e.g., reading time or eye-tracking) (Kuribayashi et al, 2021). This suggests that even as models improve at the tasks they are designed to do (e.g., masked word prediction), they may continue to diverge from humans in important, cognitively relevant ways. Additionally, as noted in *Supplementary Analysis 1*, we did replicate our analyses with ELMo, another language model

(Peters et al, 2018) that typically underperforms BERT on Word Sense Disambiguation tasks (Wiedemann et al, 2019), and with BERT-large, which contains twice as many layers and many more parameters. Interestingly, although BERT-large produced better predictions of human relatedness judgments, it was a worse predictor (as measured by AIC) of reaction time than BERT-base. This reinforces the point that improvement on one task does not entail improvement across the board—so it is no guarantee that future language models will acquire humanlike sense knowledge in the absence of other methodological interventions intended to render them more humanlike (i.e., fine-tuning them to a Word Sense Disambiguation dataset, or incorporating grounding into their training regime). Further, in *Supplementary Analysis 3*, we asked whether a different metric (Surprisal) explained more variance than the pre-registered measure (Cosine Distance); while Surprisal was indeed predictive of behavior, it did not eliminate the explanatory power of sense boundaries, consistent with the predictions of Hybrid Meaning Theory. Finally, in the Computational Modeling section, we explored several bottom-up transformations to Cosine Distance, all of which suggest that distributional statistics alone are insufficient to account for the category effects of sense boundaries.

On this note, it is worth reiterating that BERT represents a particular implementation of Pure Exemplar Theory—i.e., one in which continuous meaning representations are derived from distributional regularities in linguistic input alone. BERT (and most other neural language models) lack extra-linguistic grounding (Lake & Murphy, 2020; Bender & Koller, 2020). Thus, any semantic knowledge that relies on extra-linguistic information (e.g., perceptual experience) will be inaccessible to BERT. While this limits BERT's predictive power, it also offers a useful inferential tool: models like BERT help establish empirical limits on how much human linguistic knowledge can be captured from distributional regularities alone. As Elman (2011) notes, a

continuous meaning space could be constituted by many different dimensions of experience, including the sensorimotor or even social associations with individual words and constructions. Accordingly, future work could make use of recent developments in grounded language models (Su et al, 2019; Zellers et al, 2021; Johns, 2021) to ask whether access to particular dimensions of sensorimotor information elicits more humanlike behavior. One possible outcome is that the variance in human behavior currently explained by sense boundaries can actually be attributed to aspects of sensorimotor or social experience uncaptured by BERT. Under one interpretation, this would salvage a version of Pure Exemplar Theory, which simply admits more dimensions of human experience into the continuous state-space.

**Is Homonymy Just "Distant" Polysemy?**

We found no evidence that homonymous meanings exhibited different priming effects than polysemous meanings. This is surprising, given the extensive evidence that the two phenomena elicit systematically different behavior on a number of tasks (Rodd et al, 2002; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Klepousniotou et al, 2012; Rodd et al, 2012; Floyd & Goldberg, 2021), and the fact that they are typically treated as distinct phenomena in theoretical linguistics and lexicography (Valera, 2020). Some past work has nevertheless acknowledged the possibility of homonymy and polysemy lying along a continuum, both in theoretical cognitive linguistics (Tuggy, 1993) and experimental psycholinguistics (Rodd et al, 2002; Brown, 2008; Klepousniotou et al, 2008). As noted in the Introduction, however, the majority of work in this area has not incorporated this notion of a continuum between homonymy and polysemy into theoretical or formal models of lexical ambiguity (with some exceptions, e.g., Rodd (2020)).

Does this mean that homonymy is simply "distant" polysemy—at least when it comes to their respective impacts on cognitive processing? There are several possible reasons why it does not. First, we might simply have failed to detect a real, non-zero difference between polysemy and homonymy (e.g., because the study was under-powered). On the other hand, while there is always a real possibility of a false negative result, we obtained null results across two large N studies ($N \geq 180$); further, a power analysis suggested that we should have had 95% power to detect an effect of the size we detected in a pilot study. Combined, this suggests that the behavioral differences in this paradigm are either nonexistent, or small enough to be of negligible theoretical interest.

Second, it is possible that our operationalization of homonymy and polysemy—i.e., determining whether two meanings were listed as separate entries in the dictionary—was somehow deficient. However, it is unclear how better to operationalize these variables. Binning according to some behavioral variable (e.g., relatedness) would impose semi-arbitrary structure on a continuous space, which is precisely the question we are attempting to address. The expertise of lexicographers for Merriam-Webster and the OED may be the closest approximation to the received expert view that can be found. Nevertheless, it is possible that another operationalization, perhaps relying on finer-grained distinctions between semantic relations (e.g., metaphor vs. metonymy), could result in the detection of behavioral differences across categories of ambiguity.

Third, our original, pre-registered analyses did not account for sense dominance (i.e., when one meaning of an ambiguous word is more frequent than another), which is known to influence ease of processing (Duffy et al, 1988; Klepousniotou et al, 2008; Blott et al, 2020). However, we counterbalanced the order of the prime and target sentences across participants.

Thus, if many of the sentence pairs contained unbalanced meanings, our results would essentially be averaging across a null or small effect (i.e., moving from a subordinate to a dominant sense) and a strong effect (i.e., moving from a dominant to subordinate sense); we believe this is unlikely to account for the failure to find a significant difference in the priming effect across polysemous and homonymous pairs. Additionally, we did run a post-hoc analysis using normed dominance judgments for different sense items only (see *Supplementary Analysis 2*). This analysis replicated the effect of dominance found in past experiments (Klepousniotou et al, 2008), as well as the main effect of Cosine Distance reported in Experiments 1-2. There was also a possible main effect of Ambiguity Type for different sense pairs only—but as noted below, this main effect could be driven by uncontrolled differences among the stimuli themselves, and is not necessarily attributable to differences in the strength of priming across homonymous and polysemous stimuli (i.e., we failed to detect a Sense Boundary x Ambiguity Type interaction in both experiments).

Finally, and perhaps most importantly, it is possible that the primed sensibility judgment task is simply not well-suited for detecting a difference between homonymy and polysemy. Our failure to detect a difference on one task does not entail that the two phenomena are not psychologically distinct in general. To make this more general claim—i.e., that homonymy is "distant" polysemy—one would need to demonstrate a null effect across a number of tasks that have provided evidence for a categorical difference between polysemy and homonymy. If, by process of elimination, each task fails to elicit behavioral differences above and beyond the continuous distance between two contexts of use, one might at last conclude that homonymy and polysemy truly do lie along a continuum; if, on the other hand, some tasks *do* continue to elicit

different behavior, that would provide deeper insight into exactly when and under what conditions this categorical distinction is cognitively and behaviorally relevant.

Here, it is worth revisiting the surprising finding that accuracy did significantly differ across sentences containing homonymous and polysemous sentences, on *both* prime and target trials. Since the size of this effect did not differ across prime and target trials, this indicates that there was no difference in the priming effect itself. There are several possible explanations for this main effect. First, it could be due to uncontrolled differences in the stimuli: perhaps the sentences containing homonyms, or the homonymous items themselves, happened to be less natural than those containing polysemes. Although we adjusted for a number of features in our analyses (e.g., frequency, length, concreteness), it is possible that we failed to account for a crucial determinant of lexical processing. Second, the effect could be driven by a theoretically meaningful difference in how homonymous and polysemous words are processed. Past work (Rodd et al, 2002; Klepousniotou, 2002) has found differences in reaction time and accuracy on isolated lexical decision tasks. If accessing the meaning of a homonym involves competition from its other, unrelated meanings (Rodd et al, 2002), then sentences containing homonyms might also be genuinely harder to process than those with polysemes, even independent of priming.

**Sense Representations in a Continuous State-Space**

The experimental results reported above support Hybrid Meaning Theory, which claims: 1) word meanings are context-dependent trajectories through a continuous state-space; and 2) these trajectories are mediated by sense representations, such that *contextual distance* is

transformed into a sense-mediated *conceptual distance*. The second claim raises a number of questions about the nature of these sense representations.

First, there are a number of distinct computational mechanisms by the use of which sense representations might mediate contextual distance. In the Introduction, we distinguished between Sense Attraction, in which tokens within a sense-cluster shrink towards their centroid, and Sense Distillation, in which within-cluster variance is removed altogether, preserving only the centroid or prototypical member. An exploratory analysis provided evidence in favor of the Sense Attraction mechanism; even considering only same sense uses, we found that response time was positively correlated with contextual distance, suggesting that some within-cluster variance is preserved. Further, we applied several functional transformations to Cosine Distance, and asked which transformation yielded improvements in predicting human behavior. We found that an additive transformation to Cosine Distance best improved a statistical model's fit; crucially, the best transformation was "top-down" and explicitly used the Sense Boundary variable, i.e., information *external* to the underlying BERT-space. This suggests that distributional regularities alone—even after applying a bottom-up transformation—are insufficient to account for the emergence of sense-like representations.

This leads to a second, related question: how and when do these sense representations emerge? Given that every context of use constitutes a slight variation in meaning, what degree—or what dimensions—of variation results in the creation of a sense boundary? Klein & Murphy (2001, pg. 279) summarize the question as follows (emphasis ours):

> "If two senses are only very subtly different, it seems unlikely that speakers will develop separate entries for them, since a single entry will suffice to specify most of the meaning

> for both. If two senses are strikingly different, then a single entry will probably be unsuccessful at representing both meanings, which will presumably lead to the formation of separate entries…**What is needed is a more specific model of what causes a sense to be separately represented, from which one could derive predictions about which uses would involve the same senses and which would involve different senses**."

One promising avenue would be to look to related research on how children acquire ambiguous words. There is some evidence that children are better able to acquire new meanings for a known wordform when those meanings are related, rather than unrelated, to its existing meanings (Floyd & Goldberg, 2021). This echoes previous findings that homonyms are challenging to learn (Casenhiser, 2005), possibly because children have a bias against assuming homophony—though more recent work (Dautriche et al, 2016) suggests that children reliably postulate homophony if the exemplars presented from each meaning are sufficiently distinct. Finally, work by Srinivasan & Snedeker (2011) suggests that children rely on a common representation for polysemous words with highly regular meaning relationships (e.g., "heavy *book*" and "popular *book*"). As the authors note, this common representation could be lexical, with rules for deriving each meaning stored with the word itself (Pustejovsky, 1995); alternatively, they might rely on more general conceptual knowledge, likely reflecting systematic conceptual relations within and across languages (Srinivasan & Rabagliati, 2015). Yet to our knowledge, it remains unknown whether and when these related meanings drift apart into distinct sense-clusters. In our task, English-speaking adults demonstrated an effect of sense boundary above and beyond the distance between two contexts of use, and there was no significant difference in the size of this effect between polysemes and homonyms—suggesting that at least in adults, polysemous meanings

manifest in distinct sense-clusters. Future work could use a similar paradigm with children, and ask at what age children begin to differentiate highly related polysemous meanings.

A second avenue would be to develop hypotheses about which dimensions of contextual variability are most likely to predict the emergence of a new sense. If word meaning is at least partially grounded in sensorimotor experience (Barsalou, 1999; Pulvermüller, 2013; Bergen, 2015), one possibility is that a new sense-cluster is generated when the associated sensorimotor profile is sufficiently distinct. For example, one meaning might be more concrete than the other, as is the case with much of conceptual metaphor (e.g., "a wooden *table*" vs. "a data *table*"). Alternatively, different contexts of use might be similarly concrete, but involve different bodily effectors, different perceptual modalities, or even different instruments. For example, "*cut* the paper" and "*cut* the hair" both typically involve scissors, whereas "*cut* the grass" often involves a lawn mower. If psychological senses are motivated by sensorimotor distinctions, then one would predict that "*cut* the paper" and "*cut* the hair" are more likely to behave as same sense items, while "*cut* the paper" and "*cut* the grass" should be more likely to behave like different sense items, all other things being equal. Similarly, the difficulty of transitioning across a sense boundary might be highest when those senses have very different sensorimotor profiles.

## Conclusion

Word meaning is highly context-sensitive and often outright ambiguous. Accordingly, mental representations of word meaning must be flexible enough to accommodate this context-sensitivity. However, traditional theoretical frameworks analogize mental representations to entries in a physical dictionary, which are static and discrete; this conceptualization is challenging to reconcile with the flexible, context-dependent nature of word meaning. We reviewed evidence supporting the Continuity of Meaning Framework, in which word meanings

are conceptualized as context-sensitive trajectories in a continuous state-space; we also introduced two "hybrid" theories, which posit discrete, psychologically real categories atop this continuous space. In two behavioral experiments using a primed sensibility paradigm, we found that human behavior was best predicted by a theory that posits both continuous, flexible meaning representations as well as discrete senses.

**Acknowledgments**

Chapter 4, in full, is under review at *Psychological Review*. Trott, Sean; Bergen, Benjamin. The dissertation author was the primary investigator and author of this paper.

# CHAPTER 5: WHY DO HUMAN LANGUAGES HAVE HOMOPHONES?

Human languages are replete with ambiguity. This is most evident in *homophony*—where two or more words sound the same, but carry distinct meanings. For example, the wordform "bark" can denote either the sound produced by a dog or the protective outer sheath of a tree trunk. Estimates of the rate of homophony in English range from 7.4% (Rodd et al, 2002) to over 15%[19] (Baayen et al, 1995). Dautriche (2015) estimates the average homophony rate across languages to be 4%, with considerable cross-linguistic variability, ranging from approximately 3% in Dutch to 15% in Japanese. The prevalence of homophony, like other kinds of ambiguity, is confounding on its face. Human languages are generally thought to be shaped by pressures for efficient, effective communication (Zipf, 1949; Gibson et al, 2019). Yet ambiguity increases both the effort required for comprehension and the likelihood of miscommunication. A comparison between human and programming languages places this into relief. Programming languages, designed for efficient and errorless communication, generally abide no ambiguity at all. Why then do human languages insist on encoding distinct messages identically? Why are homophones so common?

Part of the answer appears to be that human comprehenders are adept at disambiguating ambiguous input using various contextual cues (Levinson, 2000; Wasow et al, 2005; Ferreira, 2008; Piantadosi et al, 2012). In the case of homophones, a wide array of cues to meaning are available, including the syntactic structures that words are embedded in (Dautriche et al, 2018), gestures that accompany speech (Holle & Gunter, 2007; Holler & Beattie, 2003; Kidd & Holler, 2009), and statistical aspects of linguistic context (Aina et al, 2019). The human capacity for

[19] Estimates of the rate of *polysemy* (wordforms with related meanings) are considerably higher: up to 80% of wordforms in English are thought to be polysemous (Rodd et al, 2002).

disambiguation thus creates a tolerant environment for ambiguous wordforms—explaining why as languages evolve, homophones might not be strictly selected against.

But might homophones also be selected for? Zipf (1949) argues that ambiguity is a design feature of any human communication system, resulting from a direct pressure for efficiency. A growing body of evidence is consistent with the claim that lexica are optimized for efficient communication between humans (Piantadosi et al, 2009; Gibson et al, 2019), from the way they carve up semantic domains (Regier et al, 2007; Kemp & Regier, 2012; Xu & Regier, 2014; Kemp et al, 2018; Zaslavsky et al, 2018) to the wordforms that they contain (Piantadosi et al, 2011; Piantadosi et al, 2012; Mahowald et al, 2018). This pressure for an efficient lexicon could result in a selective bias for wordforms that are particularly easy to produce and comprehend, where *ease* reflects properties such as a word's length, phonotactic plausibility, and frequency. Combined with a tolerance for ambiguity, a bias for easy wordforms could exert a pressure on lexica to "recycle" particularly optimal wordforms for multiple meanings. This pressure, termed "unification" by Zipf (1949), would increase efficiency by reducing the number of unique wordforms that speakers need to learn and encode. Furthermore, by preferentially re-using the most optimal wordforms, such a lexicon would arguably involve less effort in speaking or writing than an unambiguous linguistic system. If such a pressure exists, it should produce concentrations of homophony in optimal regions of phonotactic space—the "easiest" wordforms should be the most ambiguous. Indeed, Piantadosi et al (2012) find that English, German, and Dutch count more homophones among wordforms that are short, frequent, and phonotactically well-formed. This finding is consistent with the idea that ambiguity arises out of a pressure for efficiency.

However, homophony could also emerge in a lexicon without being directly selected for, as an indirect consequence of other factors affecting how words are distributed in a lexicon. Two indirect mechanisms could also partially (or even fully) account for the uneven distribution of homophony across a lexicon.

First, the *proportion of occupied phonotactic space* (i.e., the ratio of actual wordforms to possible wordforms) for English and every other language we are aware of will always be higher for shorter wordforms than for longer wordforms. This is because the number of possible wordforms of a given length grows exponentially with each added syllable. If a language's phonotactics permit $n$ unique syllables, then there are $n$ possible monosyllabic wordforms, approximately $n^2$ possible bisyllabic wordforms, approximately $n^3$ possible trisyllabic wordforms, and so on. In contrast, the number of actual wordforms does not grow exponentially with word length (e.g., the CELEX set of English lemmas contains approximately 7706 monosyllabic words, 15247 disyllabic words, and 11379 trisyllabic words). This means that the proportion of occupied phonotactic space will always be greater among short wordforms than long wordforms. Thus, even if words were randomly added to a lexicon, homophony would by chance be more likely to occur among short wordforms than long wordforms.

Second, the existence of *phonotactic constraints* results in a lexicon that is not uniformly distributed across the space of possible wordforms. All languages appear to impose idiosyncratic constraints on sounds and their combinations—for example, English does not allow the velar nasal */ŋ/* in syllable onsets, unlike Vietnamese; but English does allow consonant clusters like */st/*, unlike Japanese. Phonotactic regularities narrow the space of possible wordforms considerably (Dautriche et al, 2017). By limiting the range of possible wordforms and biasing the formation and evolution of the lexicon, these phonotactic constraints could also increase the

prevalence of homophones. Critically, they could do so even without a direct pressure to reuse entire wordforms. Even a pressure to merely statistically reuse certain phonological sequences more often would increase the likelihood of homophones overall, and particularly among the most phonotactically probable wordforms.

Both of these mechanisms offer indirect causal pathways whereby a drive for efficiency could lead to increased homophony. For example, more phonotactically regular words could be easier to learn (Jusczyk et al, 1994; Gathercole et al, 1991; Munson, 2001; Coady & Aslin, 2004), which would lead to more phonotactically probable words being more likely to be transmitted across generations, or less phonotactically words becoming more phonotactically probable through imperfect intergenerational transmission. This in turn could result in increased homophony, particularly among highly probable wordforms. Once again, though, both phonotactics and the distribution of word lengths in a lexicon could in principle lead to the emergence of homophony without a direct, selective pressure for the preferential reuse of specific, optimal wordforms (as hypothesized by Zipf, 1949). Furthermore, both factors should be most likely to produce homophones in exactly those regions of phonotactic space reported by Piantadosi et al (2012): among short, phonotactically plausible wordforms.

It is currently unknown, however, how much homophony exists due to these simple, distributional characteristics of languages alone. As a consequence, no evidence exists for or against an efficiency-motivated direct pressure for homophony, as hypothesized by Zipf. The current work asks two primary questions. First, to what extent is the **amount** of homophony found in real human lexica attributable to indirect and uncontroversial factors such as length and phonotactic regularities, without a direct pressure to reuse existing wordforms? And second, to

what extent are these indirect factors responsible for the **concentration** of homophony within optimal regions of the lexicon?

To answer these questions, we constructed five series of artificial lexica, designed to mirror the phonotactic regularities and word lengths of the real lexica of English, Dutch, German, French, and Japanese. The generative model was an adaptation of the model used by Dautriche et al (2017), in which a language's phonotactics were learned by training an *n*-phone Markov Model on the set of unique wordforms in a lexicon. By observing the patterns of sounds and sound combinations in a language, such a model can learn to encode phonotactic rules about which sounds a word can start and end with, which sounds can occur in what sequence, and so on. For each language, this model was then used to generate 10 artificial lexica, all matched for the total number of words as well as the distribution of word lengths. For example, if the real lexicon has 5,000 monosyllabic words, then each of the artificial lexica will also have 5,000 monosyllabic words. Furthermore, the distribution of sounds within and across those words will approximate the phonotactics of the real language. These artificial lexica had no constraints regarding homophones, reflecting a general tolerance for ambiguity; however, they also did not contain a parameter biasing them *toward* the reuse of existing wordforms. Each artificial lexicon thus represents one answer to the questions: 1) **how much** homophony can be expected to emerge in a lexicon as a function of just the real, observed phonotactic regularities and the real, observed distribution of word lengths; and 2) where should we expect to find the largest **concentrations** of homophony as a function of these factors? They thus serve as a baseline characterization of the effects of indirect causes of homophony. Comparing the real lexica to these artificial ones reveals how much more or less homophony the real languages display—and

how much more or less concentrated it is—than would be expected without any direct pressure for or against homophony[20].

Note that these artificial lexica are not intended to serve as plausible models of lexicon formation and change. Rather, as described above, they serve as statistical baselines in the attempt to understand which theoretical parameters are necessary to explain the existence and distribution of homophony in real lexica. For this reason, the artificial lexica are parameterized solely by each particular language's phonotactics and distribution of word lengths.

The data and code to reproduce these analyses can be found on GitHub (https://github.com/seantrott/homophone_simulations).

## Current Work

### Materials and Methods

**Data.** The English, German, and Dutch lexica were sourced from the CELEX lexical database (Baayen et al, 1995). For French, we used the French Lexique (New et al, 2004). For Japanese, we used the Japanese CallHome Lexicon (Kobayashi et al, 1996). We restricted our analysis to lemma-only forms. Additionally, following Piantadosi et al (2012), we also excluded any words containing spaces, hyphens, or apostrophes. This resulted in 41,887 entries for English (with 35,107 unique phonological forms), 51,719 entries for German (with 50,435 unique phonological forms), 67,477 entries for Dutch (with 65,260 unique phonological forms), 47,782 entries for French (with 37,278 unique phonological forms), and 51,147 entries for Japanese (with 40,449

[20] Note that our statistical models do not include a measure of frequency, even though this is included in the original model built in Piantadosi et al (2012). This is because it would not be meaningful to estimate frequency for the words in the artificial lexica.

unique phonological forms). As in Piantadosi et al (2012), words with multiple parts of speech were counted as homophones[21].

**Methods.**

**Estimating number of syllables.** Our primary determinant of word length was Number of Syllables (or Number of Morae, in the case of Japanese; see below). While the real lexica annotated this information for each lexical entry, it had to be estimated for the artificial lexica. To ensure a fair comparison, we applied the same estimation procedure to wordforms in the real lexica and wordforms in the artificial lexica.

For English, Dutch, German, and French, Number of Syllables was estimated by counting the number of vowels occurring in a wordform's phonetic transcription. The set of possible vowel characters for a given language was transcribed by hand and can be found on the project's GitHub page.[22]

Since Japanese has been characterized as a mora-timed, rather than syllable-timed language (Port et al., 1987), we calculated Number of Morae instead of Number of Syllables. In addition to counting the number of vowels in a Japanese wordform, we counted the number of nasal codas, as well geminate consonants (e.g., "kk*"* in *Hokkaido*, or *"*gg" in *doggu*). It should be noted that the results we report below—both the replication of Piantadosi et al (2012), and the comparison to the artificial lexica—are qualitatively similar whether word length in Japanese is estimated using Number of Syllables or Number of Morae.

**Counting number of homophones.** Following Piantadosi et al (2012), we defined Number of Homophones as the number of lexical entries with an identical phonological form as

[21] Importantly, this should only serve to *inflate* the estimated amount of homophony in naturally-occurring languages relative to the amount of homophony in the artificial lexica. Thus, it would actually work against the effects reported below (i.e., the artificial lexica exhibiting more homophony than the real lexica).
[22] Link: https://github.com/seantrott/homophone_simulations

some target entry.[23] This means the smallest possible value for Number of Homophones would be 0 (i.e., there are no other words with the same form in a given lexicon), and the largest possible value would be one less than the size of the lexicon (i.e., all words share the same form).

After identifying the number of homophones for each entry in a lexicon, we reduced each lexicon to the set of unique phonological wordforms (e.g., the 41,887 entries in English were reduced to 35,107 unique forms).

**Building the phonotactic model.** In order to estimate the phonotactic plausibility of wordforms in a lexicon, as well as to generate phonotactically plausible novel wordforms (see below), it was first necessary to model the phonotactics of each language. We adapted the procedure used in Dautriche et al (2017)[24], which is described briefly below.

The phonotactics of a target language can be learned by observing, for all wordforms in that language, which phonemes appear in what position and in what sequence. Specifically, an *n*-phone model calculates the probability of observing some phoneme in position *i* given the previous *n-1* phonemes. For example, a 2-phone (biphone) model would condition the probability of observing some phoneme as a function of the previous phoneme, i.e., $p(X_i \mid X_{i-1})$. We included special symbols for the START and END of a word so that the model would also learn which phonemes are most likely to begin and end a word in a given language. Note that unlike Piantadosi et al (2012), these models were trained using the set of unique *types* (i.e., wordforms), rather than *tokens* (i.e., the actual instances of each wordform); this is because

[23] As pointed out by an anonymous reviewer, it is possible that the lexical resources we used, including CELEX, count as homophony some meanings that are actually polysemous. If this is the case, our estimates of homophony should actually be inflated for the real lexica, which would work against the effects reported below (i.e., the artificial lexica displaying higher incidences of homophony overall).

[24] Link to GitHub associated with Dautriche et al (2017): https://github.com/SbllDtrch/NullLexicons

training on tokens conflates phonotactic probability with frequency. This is analogous to the main approach taken in Dautriche et al (2017).

While previous work (Dautriche et al, 2017) found that a 5-phone model effectively captured phonotactic dependencies in English, Dutch, German, and French, we sought to independently determine the optimal *n* for each language, particularly because Japanese has notably shorter syllables than the other four languages. To do this, we followed a similar procedure as reported in Dautriche et al (2017) and Futtrell et al (2017). For each real lexicon, we first extracted the set of unique wordforms (e.g., 35,107 wordforms in English), then performed a series of train/test splits (75% train, 25% test). For each split, we trained a series of *n*-phone models ranging from $n$=1 to $n$=6 on the wordforms in the training set, then evaluated the probability of wordforms in the held-out test set. The basic motivation for this approach is as follows: the optimal *n*-phone model for a language's phonotactics should be the model that, when trained on a set of real wordforms, maximizes the probability of held-out wordforms that also appear in that lexicon. Following Futtrell et al (2017), we ran a series of one-tailed two-sample t-tests on the set of log-likelihoods of held-out wordforms obtained from each successive *n*-phone model—i.e., the log-likelihoods obtained from the 2-phone model were compared to the 1-phone model, those from the 3-phone model were compared to the 2-phone model, and so on. The optimal *n* for a given lexicon was the smallest *n* that represented a significant improvement over the *n-1* model for the same set of wordforms. Note that $\log_{10}$ was used to calculate log likelihoods (and subsequently, surprisal); the results are not qualitatively different when using $log_2$ instead.

The mean log-likelihood calculated for held-out wordforms in each language are visualized in *Figure 18* below. Critically, we found that for English, Dutch, and German, the 5-

phone model represented a significant improvement over the 4-phone model. That is, held-out wordforms were significantly more likely under the 5-phone model than the 4-phone model for English ($t = 4.05, p < .001$), Dutch ($t = 3.55, p < .001$), and German ($t = 7.31, p < .001$). However, the 6-phone model either did not improve or actually decreased model fit (suggesting overfitting) in each language (all $t \leq 0$). The 4-phone model was optimal for French ($t = 8.67, p < .001$) and Japanese ($t = 4.08, p < .001$). Thus, a 5-phone model was used to evaluate the probabilities of wordforms in English, Dutch, and German (and generate artificial lexica for those languages), and a 4-phone model was used for French and Japanese.

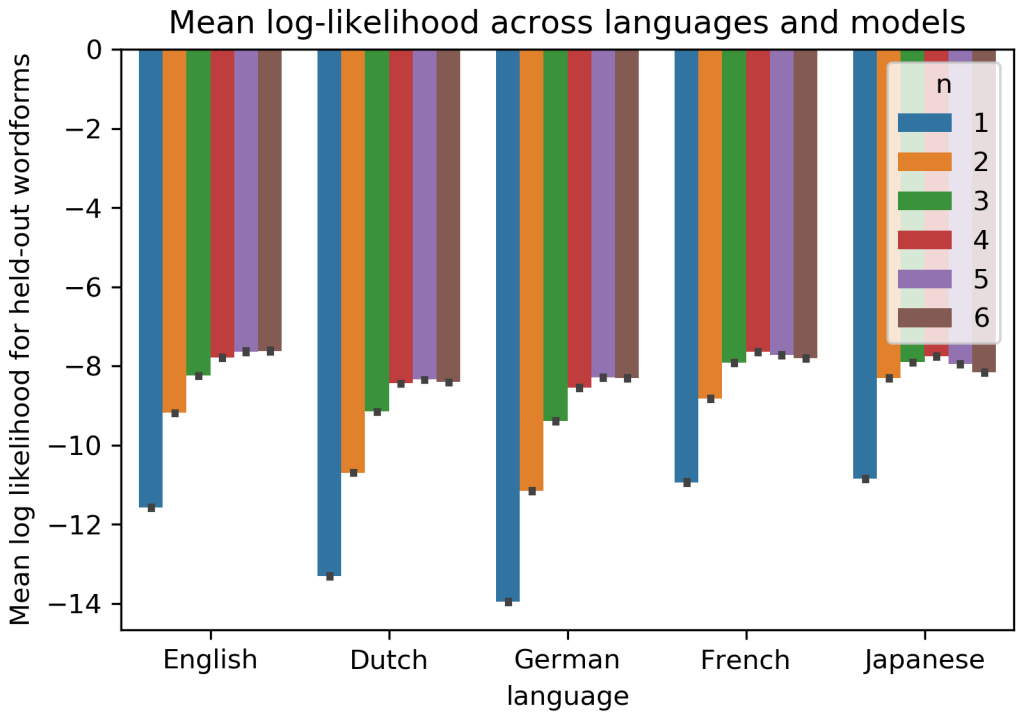


Figure 18: Mean log-likelihood of held-out wordforms for each n-phone model, across languages. Higher values (i.e., less negative) indicate higher probability under that model. For English, Dutch, and German, increasing n up to 5 significantly improved model fit over the 4-phone model; a 6-phone model did not improve fit. For French and Japanese, a 4-phone model was the highest n representing an improvement over the (n-1) model.

This model allows us to evaluate the probability of a given wordform, which can be defined as the product of all the transitional probabilities between each phoneme in that wordform (including the start and end symbols). The Surprisal of a given wordform is thus defined as the negative log probability of observing that particular sequence of phonemes: Surprisal(word) = -log(p(word)). As in Piantadosi et al (2012), we normalized this measure to the number of phonemes in a word to ensure that surprisal could be compared across words of different length: Normalized Surprisal = Surprisal(word) / Length(word).

Once the model was built for each language, it was then used to generate novel wordforms in an iterative manner. For each word, the model began with the START symbol, then generated a phoneme conditioned on that start symbol (i.e., one of the phonemes likely to occur at the beginning of the word). The next phoneme was then conditioned on the first phoneme and the START symbol, and so on, until the model produced the END symbol, signaling the end of the word.

Finally, as in Dautriche et al (2017), we assigned non-zero probability to unobserved phoneme sequences using an identical smoothing procedure; they report that "optimal smoothing was obtained with Laplace smoothing with parameter .01" (pg. 132), so this was the value we used in configuring the phonotactic model.

**Generating artificial lexica.** We generated 10 artificial lexica for each real lexicon. First, we identified the number of words in the real lexicon, as well as the distribution of their lengths, as measured by Number of Syllables (see above for the estimation procedure). Each artificial lexicon was constrained to have the same overall number of *words* (not *wordforms*) as the corresponding real lexicon, as well as the same distribution of word lengths. For example,

since the real English lexicon has 7706 monosyllabic words, each artificial English lexicon was also constrained to have 7706 monosyllabic words.

We then built a phonotactic model for the real lexicon as described above, and used this model to generate wordforms for each artificial lexicon. For each potential wordform, we estimated the Number of Syllables to determine whether to add it to the artificial lexicon—e.g., if the word had 1 syllable and the artificial lexicon still had fewer monosyllabic words than the real lexicon, the word was added to the lexicon; otherwise, it was discarded. No other constraints were placed on the generation of wordforms; we allowed the model to generate real wordforms, as well as wordforms that were homophonous with wordforms already in the lexicon. This process continued until the artificial lexicon had the same number of words of each length as the real lexicon.

Note that the models used to generate the artificial lexica were trained on the entire set of unique wordforms for the target lexicon; however, qualitatively similar results were obtained using a 50/50 split of the target lexicon to generate and evaluate wordform phonotactic probability (see *Supplementary Analysis 5*).

## Results

### Replication and extension of previous findings.

First, we replicated the primary analysis reported by Piantadosi et al (2012) on the real lexica of English, Dutch, and German, and extended this analysis to two non-Germanic languages: French and Japanese. Using a Poisson regression, we asked whether a wordform's #Homophones (the number of additional, distinct meanings) was related to its length in syllables (#Syllables) and its phonotactic plausibility (Surprisal). As in Piantadosi et al (2012), we used

the Normalized Surprisal measure described above, obtained by dividing a wordform's Surprisal by its length in phones.

We found significant, negative relationships in the real lexica between #Homophones and #Syllables (or #Morae[25] in Japanese) for English [$\beta = -0.72$, SE = 0.03, $p < .001$], German [$\beta = -0.69$, SE = .04, $p < .001$], Dutch [$\beta = -1.11$, SE = 0.03, $p < .001$], French [$\beta = -0.35$, SE = 0.02, $p < .001$], and Japanese [$\beta = -1.01$, SE = 0.01, $p < .001$]. That is, for all five languages, shorter wordforms were more likely to have more homophones—consistent with the notion that lexica recycle short wordforms for multiple meanings.

However, we found positive[26] relationships between Normalized Surprisal and #Homophones across all real languages but Japanese, i.e., *less* phonotactically plausible wordforms (as measured by a 5-phone model or 4-phone model, as appropriate) were more likely to have more homophones. This was true for English [$\beta = 0.78$, SE = 0.03, $p < .001$], German [$\beta = 0.86$, SE = 0.06, $p < .001$], Dutch [$\beta = 0.997$, SE = 0.04, $p < .001$], French [$\beta = 0.73$, SE = 0.04, $p < .001$], but not Japanese [$\beta = 0.0004$, SE = 0.031, $p = .99$]. This is in contrast to the original result reported by Piantadosi et al (2012), who found a negative relationship between Normalized Surprisal and #Homophones in German and Dutch.

There are several possible explanations for the disparity between our results and those of Piantadosi et al (2012). First, while Piantadosi et al (2012) used a 3-phone model to determine phonotactic plausibility, we used 4-phone and 5-phone models to estimate wordform probability,

[25] Like syllables, a mora is a unit of timing, and is usually considered the basis of the sound system in Japanese. A single mora in Japanese is constituted by a vowel (or an onset and a vowel); nasal codas also constitute a separate mora, as does the first part of a geminate consonant.

[26] Note that negative relationships were obtained between the non-normalized Surprisal measure and Number of Homophones across each language; these results are described in Supplementary Analysis 2. However, this non-normalized Surprisal measure conflates phonotactic plausibility with word length, which is why Normalized Surprisal may be a better measure overall.

which were found to improve model fit over a 3-phone model (see *Figure 18*). Second, our models were trained using lexical types, as opposed to tokens (which would conflate frequency with phonotactic probability). And third, our estimates were not calculated using held-out wordforms, as they were in Piantadosi et al (2012). This final explanation is explored in *Supplementary Analysis 3*; using 10-fold cross-validation to obtain our surprisal estimates, we found that the coefficients for Normalized Surprisal were closer to 0 for all the real lexica, and negative in Japanese. Thus, a likely reason for the disparity is that the surprisal estimates given here were not calculated using held-out wordforms.

However, the central question of the current work concerns the comparison between the real and artificial lexica. The results of these comparisons are described in detail below, both concerning the **amount** of homophony across the real and artificial lexica, as well as where those homophones are **concentrated**.

**Simulated lexica exhibit higher upper-bounds on homophony.**

We operationalized the **amount** of homophony in three ways. First, we measured the Maximum Number of Homophones per wordform—that is, in a given lexicon, how many homophones does the most homophonous wordform have? Second, we measured the Mean Number of Homophones per wordform. And third, we measured Homophony Rate: how many wordforms in a lexicon have *at least* 1 homophone? In all cases, more positive values reflect a greater amount of homophony. For each measure in each language, we compared the distribution of values obtained from the simulated lexica to the value in the real lexicon. This enabled us to ask the question: to what extent can the **amount** of homophony in a language be attributed to a selective pressure for lexical ambiguity, as opposed to an emergent outcome of a language's phonotactics and distribution of word lengths? Note that for all of these measures, the values

obtained for the real and artificial lexica were significantly different[27] ($p < .001$), except where noted otherwise.

Across all five languages, the simulated lexica had a significantly larger Maximum Number of Homophones on a single wordform (see *Figure 19* below). For example, the most homophonous wordforms in the real English lexicon had at most 7 homophones, while the most homophonous wordforms in the simulated English lexica had anywhere from 17 to 28 homophones (M = 19.8, SD = 3.3). This difference was particularly stark for Dutch: the most homophonous wordform in the Dutch lexicon had 5 homophones, while the maximum number of homophones per wordform in the simulated lexica ranged from 72 to 116 (M = 97.1, SD = 15.13).

[27] Significance was determined by comparing a given test statistic for the real lexicon $t_{real}$ to the corresponding distribution of test statistics obtained from the artificial lexica, $T_{artificial}$. Each of these values was centered according to the mean of $T_{artificial}$, denoted here as $T'_{artificial}$ and $t'_{real}$. We then conducted a two-tailed significance test, i.e., calculating the proportion of values in $|T'_{artificial}|$ that were greater than or equal to $|t'_{real}|$. This proportion corresponds to a p-value; e.g., if all the values in $|T'_{artificial}|$ are less than $|t'_{real}|$, $p = 0$.

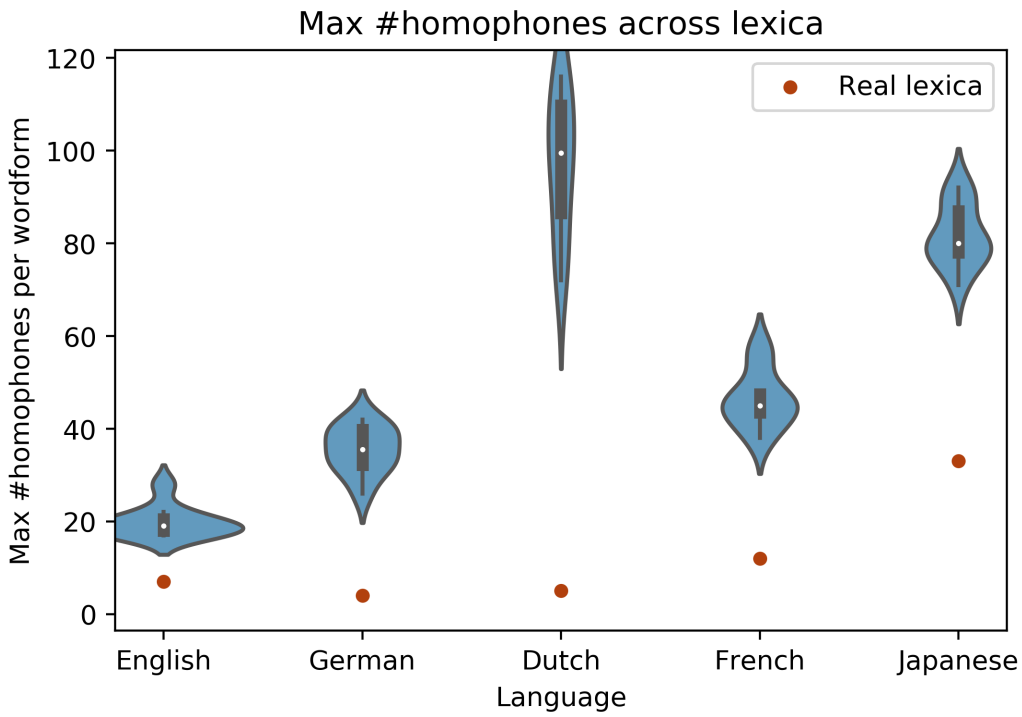


Figure 19: For each language, the most homophonous wordforms in the artificial lexica (shown by the violin plots) have more homophones than the most homophonous wordforms in the real lexica (shown by the orange dots). The artificial lexica uniformly exhibit a higher upper-bound (Maximum Number of Homophones) on homophony.

As expected, there was considerable variability across the five languages in how much homophony was tolerated per wordform. For example, the real Japanese lexicon exhibited a much higher upper-bound on homophony (33) than the real German lexicon (4); this is not surprising, given the limited syllable inventory of Japanese (on the order of 100 possible syllables) relative to German (over 1000 possible syllables, conservatively). Importantly, however, the simulated Japanese lexica still had more homophones per wordform than their real counterpart, ranging from 71 to 92 (M = 81.6, SD = 6.67). In other words, despite inter-linguistic variability, the simulated lexica in each language all exhibited higher upper-bounds on how

much homophony was tolerated for a given wordform—the most homophonous wordforms were considerably more ambiguous, sometimes by an order of magnitude (e.g., in Dutch).

Similarly, with the exception of Japanese ($p = .5$), wordforms in the simulated lexica had a significantly larger Mean Number of Homophones than wordforms from their real counterparts (see *Figure 20* below for an illustration); in Japanese, the Mean Number of Homophones per wordform was at least as high in the artificial lexica as it was in the real lexicon.[28] For example, wordforms in English have on average 0.19 homophones; in contrast, the average number of homophones per wordform in the simulated English lexica ranged from 0.22 to 0.23 (M = 0.22, SD = 0.003). Again, there was considerable inter-linguistic variability; wordforms in the real Japanese lexicon have more homophones on average (0.26) than wordforms in the real German lexicon (0.02). However, in each language, the average number of homophones per wordform was at least as large in the simulated lexica as the real counterparts—and for four of the five languages, wordforms in the simulated lexica were, on average, *more* ambiguous than those in the real lexica.

[28] Note that for Japanese, the Mean Number of Homophones per wordform is actually higher in the artificial lexica than the real lexicon with the use of a 5-phone model, rather than a 4-phone model.

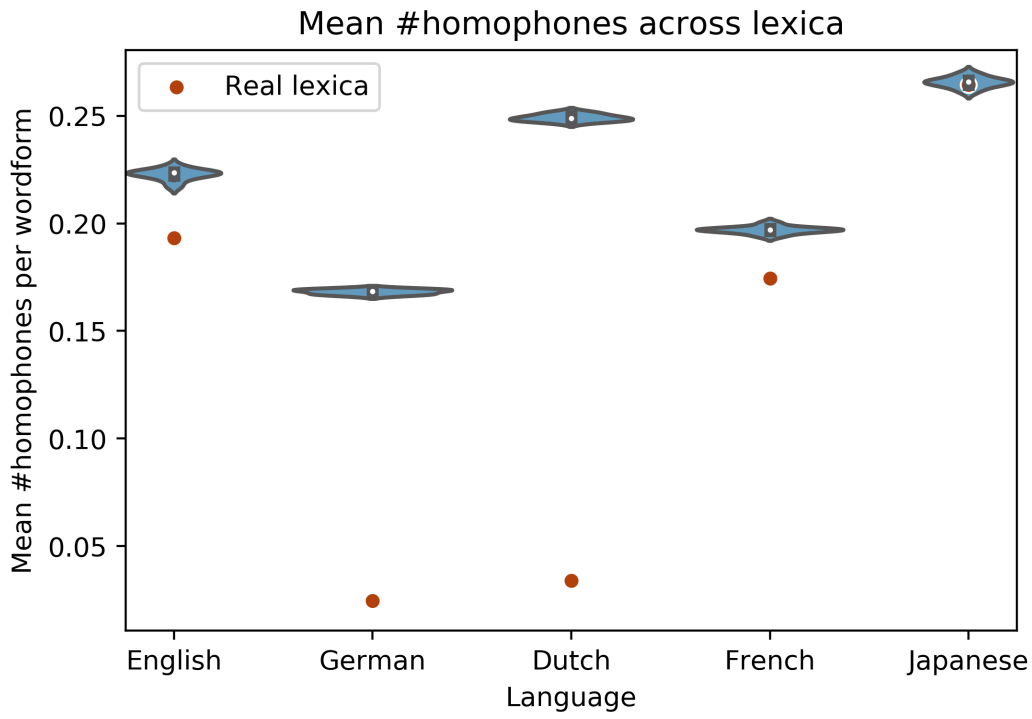


Figure 20: In every language but Japanese, wordforms in the artificial lexica (shown by violin plots) have more homophones (Mean Number of Homophones) on average than wordforms in the real lexica (shown by orange dots). In Japanese, the Mean Number of Homophones per wordform is at least as high in the artificial lexica (M = 0.27, SD = 0.002) as the real lexica (.26).

The results for the Homophony Rate (i.e., the proportion of wordforms with at least one homophone) across real and simulated lexica were more mixed (see *Figure 21* below).

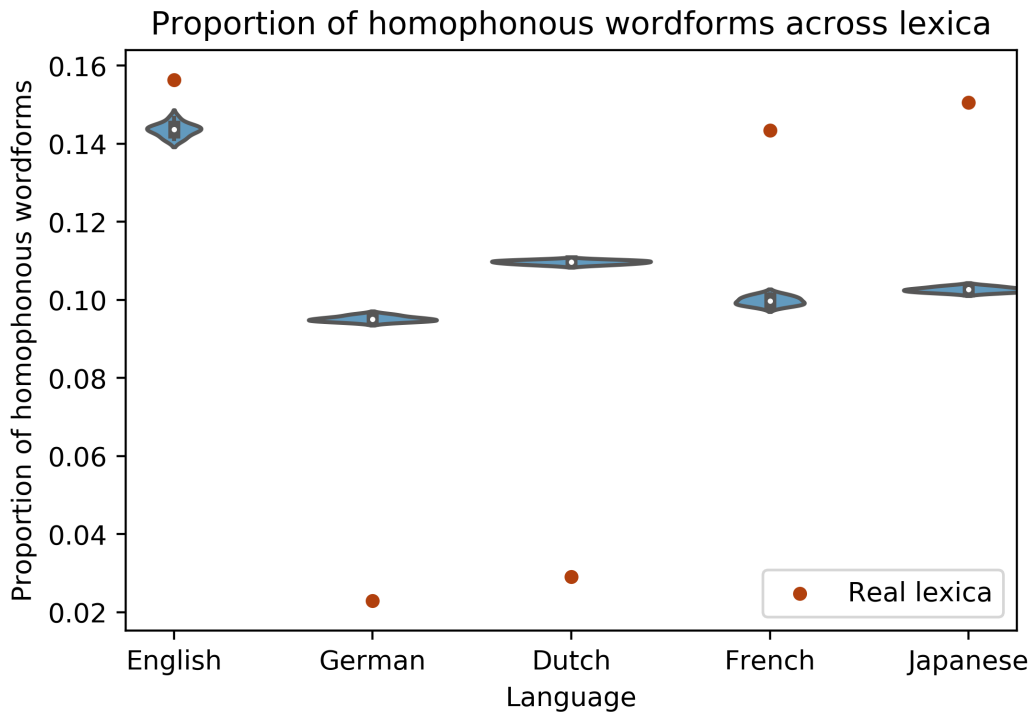


Figure 21: The artificial Dutch and German have a higher proportion of wordforms with at least one homophone (shown by the violin plots) than their real counterparts (shown by the orange dots). However, the artificial French, Japanese, and English artificial lexica have lower Homophony Rates than the real lexica.

In two languages (German and Dutch), the simulated lexica had significantly more homophonous wordforms, sometimes by a factor of 2x or 3x; for example, the homophony rate in the real Dutch lexicon was 0.03, while the rate in the simulated lexica ranged from 0.108 to 0.11 (M = .11, SD = .0004). On the other hand, the Homophony Rate in the real English lexicon (0.156) was significantly higher than the rate the simulated lexica (M = 0.143, SD = .002); similarly, the Homophony Rate for the real French and Japanese lexica were significantly higher than that for the artificial lexica.

Together, these results suggest that the **amount** of homophony in the five real lexica is not the result of a direct pressure for ambiguity. In fact, the real lexica actually display *less* homophony than the artificial ones in some measures, particularly the upper-bound of homophones tolerated for a given wordform and the mean number of homophones per wordform. This means that merely the pressure for highly probable phonotactic sequences, combined with the observed distribution of word lengths, can produce concentrations of homophony in a lexicon that are as dense or denser than in real lexica, without a direct pressure to recycle entire wordforms.

**Simulated lexica exhibit more efficient reuse of optimal wordforms.**

We then asked whether homophones were more concentrated in optimal regions of phonotactic space in the simulated lexica or their real counterparts. That is, to what extent do the phonotactics of a language, as well as its distribution of word lengths, account for the finding that more optimal wordforms tend to have more homophones?

In order to assess the degree to which homophony was optimally distributed in a lexicon, we regressed a wordform's #Homophones against two operationalizations of wordform optimality: its length (#Syllables) and its phonotactic plausibility (Normalized Surprisal). For each lexicon, we extracted the following information from the model: 1) pseudo-$R^2$, as a measure of overall model fit; 2) the coefficient for #Syllables; and 3) the coefficient for Normalized Surprisal. A larger, more positive value for (1) reflects more efficient reuse overall, and more negative values for (2) and (3) reflect more efficient reuse along those particular dimensions of wordform optimality. Then, for each language, we compared each of these test statistics from the real lexicon to the distribution of test statistics obtained from the corresponding simulated lexica. The significance for each of these comparisons was assessed in the same way as above. All of

the comparisons described revealed significant difference. To preview the overall finding, in all cases, the simulated lexica exhibited *stronger* effects (i.e., more optimally distributed wordforms) than their real counterparts.

Across all five languages, the distribution of pseudo-$R^2$ values obtained from the simulated lexica were significantly higher than the pseudo-$R^2$ value from the real lexicon (see *Figure 22* below). Pseudo-$R^2$ reflects a model's goodness-of-fit, i.e., how well the predictors in a model explain variance in the dependent variable. Thus, this indicates that two operationalizations of wordform optimality—its length, and its phonotactic plausibility—were better predictors of homophony across all of the simulated lexica than their real counterparts, for each language. For example, the pseudo-$R^2$ for the model constructed on the real English lexicon was .143, while the mean for the simulated lexica was 0.17 (SD = .004). Some differences were even starker: the pseudo-$R^2$ for the real German lexicon was .09, while the distribution of pseudo-$R^2$ values for the simulated German lexica averaged more than twice that (M = 0.231, SD = .003). Concretely, this means that homophony is better predicted by wordform optimality in the artificial than real lexica.

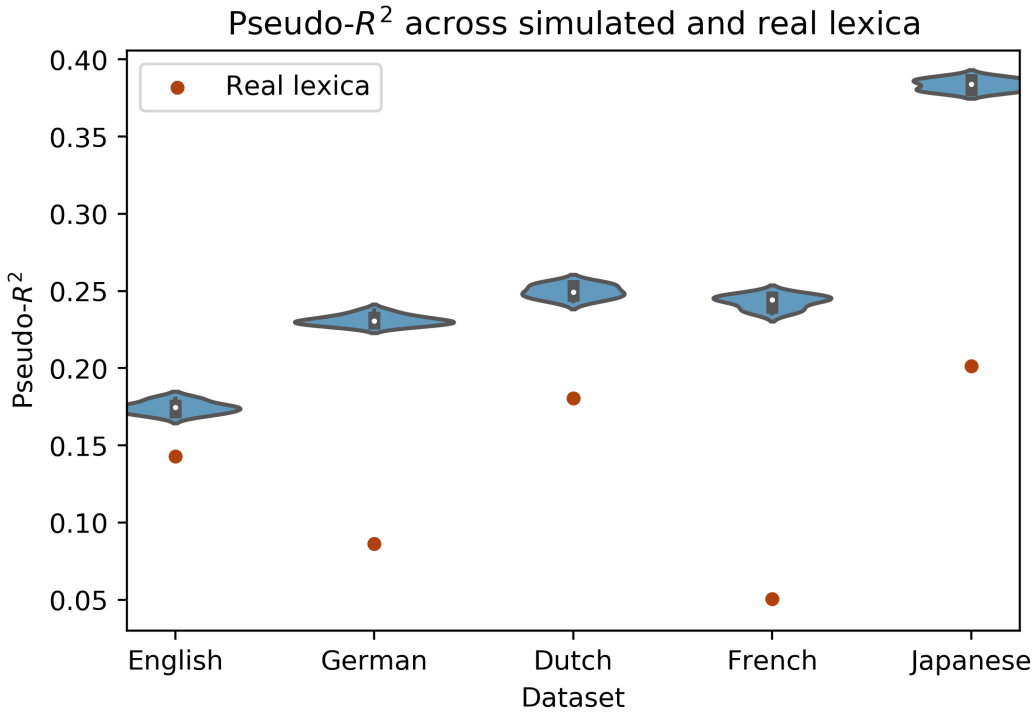


Figure 22: We built a series of Poisson regression models predicting #Homophones from #Syllables and Normalized Surprisal. In each language, the models constructed for the artificial lexica (shown by violin plots) exhibit better model fit (larger pseudo-$R^2$) than the models constructed for the real lexicon (shown by orange dots).

Further evidence comes from direct comparison of the coefficients for both predictors (Number of Syllables and Surprisal) across the real and artificial lexica. As reported earlier, the real lexica all exhibited negative relationships between Number of Syllables and Number of Homophones—i.e., short wordforms have more homophones in all five languages. However, the simulated lexica exhibited significantly stronger relationships, as depicted in *Figure 23* below.

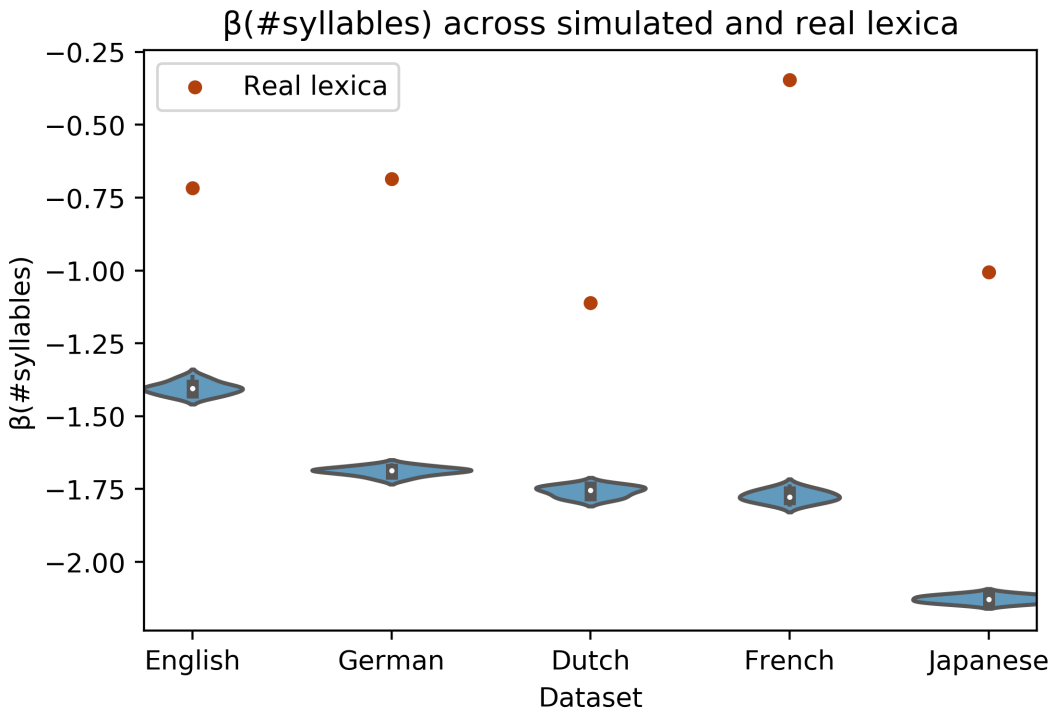


Figure 23: Word length (as measured in #Syllables) is a better predictor of homophony in the artificial lexica (shown by the violin plots) than the real lexica (shown by orange dots).

For example, the coefficient for Number of Syllables in the real English lexicon was -0.717, but the coefficients for the simulated English lexica were approximately twice as large (M = -1.4, SD = .02). In some cases, the difference was even larger, as in French: here, the coefficients for the simulated lexica (M = -1.77, SD = .02) were approximately five times as large as the coefficient for the real lexicon (-.35).

Even more striking results were obtained for Surprisal: the real lexica actually exhibited positive relationships between Surprisal and Number of Homophones, while the artificial lexica all demonstrated negative relationships (see *Figure 24*); these differences were significant for each language. In other words, the artificial lexica reused short, phonotactically plausible wordforms to a greater extent than did their real counterparts.

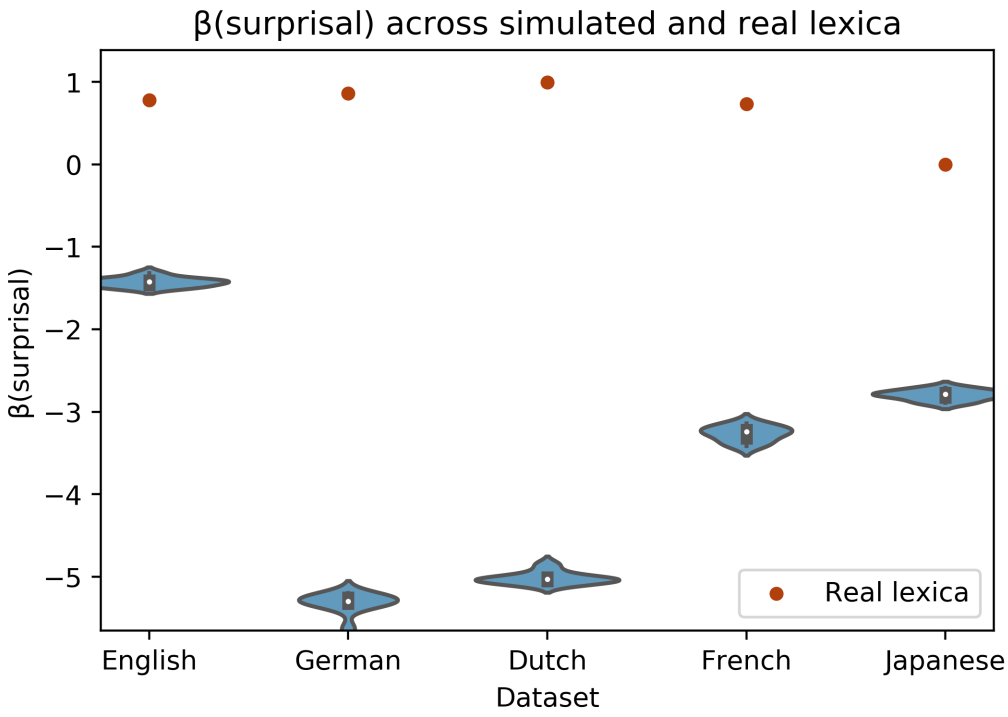


Figure 24: Phonotactic Surprisal was more negatively correlated with Number of Homophones (i.e., more probable wordforms had comparatively more homophones) in the artificial lexica (shown by violin plots) than real lexica (shown by orange dots).

## General Discussion

In the current work, we asked whether the prevalence of homophony across five languages—English, German, Dutch, French, and Japanese—could be plausibly attributed to a direct pressure to recycle optimal wordforms. We reasoned that even without a direct pressure *for* ambiguity, an absence of a pressure *against* ambiguity should result in some amount of homophony in a lexicon, simply as a result of a language's phonotactic constraints and the distribution of words across different lengths. Under this view, the selective pressure is for well-formed phonotactic sequences as opposed to entire wordforms; the pressure to use well-formed sequences could

result in homophony, particularly for the most phonotactically probable wordforms. Furthermore, given that the proportion of occupied phonotactic space will always be highest for short wordforms, homophony should also be most likely to occur in short words.

We tested this view by simulating a series of artificial lexica for each of the five languages. Across all five languages, we found that wordforms in the real lexica had either fewer or an equivalent number of homophones on average as wordforms in the artificial lexica (in every language but Japanese, wordforms in the artificial lexica were *more* ambiguous on average than those in the real lexica. The real lexica also uniformly exhibited lower upper-bounds on the number of homophones tolerated per wordform. This was true despite considerable cross-linguistic variability in the propensity towards homophony overall (e.g., Japanese vs. Dutch); in each language, the artificial lexica surpassed their real counterparts in terms of the degree to which a wordform could be saturated with many meanings. The main exception to this trend was Homophony Rate (the proportion of wordforms with at least one homophone): for English, French, and Japanese (but not German and Dutch), the real lexica had higher Homophony Rates than the artificial lexica. This will be discussed in more detail below. Finally, statistical analyses of *where* these homophones were distributed revealed that homophones in the real lexica were concentrated less efficiently in "optimal" regions of phonotactic space: across all languages, word length and phonotactic plausibility—taken as operationalizations of wordform optimality—were better predictors of homophony in the artificial lexica than the real lexica (see *Figures 5-7*).

There are two conclusions to be drawn from these results. First, neither the amount of homophony in these five real languages, nor the apparent concentration of homophones among optimal regions of phonotactic space, requires explanation by a direct pressure to recycle entire wordforms. Rather, homophony appears to be a natural and perhaps inevitable consequence of

other features of a language—i.e., its phonotactics and distribution of word lengths. Of course, these features may themselves be related to efficiency, as noted in the Introduction—but indirectly so.

Second, real lexica may actually be the product of a pressure *against* homophony. The artificial lexica were modeled using only two parameters: the phonotactics of the target lexicon and a particular distribution of word lengths. They were not designed to explicitly select *for* homophony, nor did they contain a parameter selecting *against* homophony. In other words, they reflect the consequence of allowing the phonotactics of a language to determine its space of realized wordforms, under the assumption that the speakers of that language place no upper limit on how many homophones are tolerated per wordform. This resulted in considerably more homophones per wordform than observed in real languages. For example, wordforms in the real Dutch lexicon had at most 5 homophones, whereas the average upper-bound in the Dutch lexica was 97—more than 16 times as high. Furthermore, homophony in the artificial lexica was more likely to be found among more optimal wordforms.

One explanation for this result is that real lexica are subject to a pressure against *oversaturating* the same wordform with too many unrelated meanings—no matter how "optimal" it is. Clearly this pressure is not absolute: homophony does still exist (to varying degrees) in real languages—and in fact, some languages (French, English, and Japanese) had a higher proportion of wordforms with *at least one* homophone than their artificial counterparts. This suggests that the pressure is not against the existence of homophony per se, but rather, could reflect a constraint on the extent to which any given wordform can be saturated with distinct, unrelated meanings. Assigning too many unrelated meanings to the same signal could impede communication or learning (Casenhiser, 2005; though see Dautriche et al, 2018), and may thus

be selected against. Such a pressure against oversaturation is roughly analogous to what others have termed *diversification* (Zipf, 1949) or a pressure for *clarity* (Piantadosi et al, 2012). However, unlike Zipf (1949), we find no opposing pressure towards *unification*; instead, homophony appears to emerge naturally as a function of other pressures (e.g., phonotactics), and is attenuated in particular wordforms (i.e., it does not reach the potential predicted by that wordform's phonotactics) due to a pressure against oversaturation.

There are a number of explanations for how this direct or indirect pressure against over-saturation might come about. For example, the attenuation of homophony could manifest as a kind of *smoothing* of high-probability phoneme sequences across phonological neighborhoods as opposed to being concentrated in a specific wordform. (A wordform's neighborhood is the set of wordforms differing from it in only one phoneme.) This could satisfy the pressure to reuse well-formed phonotactic sequences while also avoiding potential impediments to communication caused by overloading the same high-probability wordform with too many meanings.

This account leads to testable predictions. If real lexica are subject to this smoothing process, they should have larger phonological neighborhoods than the artificial lexica, which were placed under no pressure against ambiguity. Indeed, previous work using an identical generative model (Dautriche et al, 2017) found exactly this: across four languages (English, German, Dutch, and French), real lexica exhibit more "clumpiness" (i.e., larger and more densely connected neighborhoods) than ought to be expected merely as a function of those languages' phonotactics. We extended a subset of their analyses to the set of artificial lexica we constructed, counting as "neighbors" any two wordforms that could be converted into each other via one phoneme substitution, deletion, or insertion (Luce & Pisoni, 1998; Vitevitch & Luce, 1999; Dell & Gordon, 2003). Under this definition of neighbor, the neighbors of the word *cat*

would include *rat* (substitution), *at* (deletion), and *cast* (insertion). Consistent with prior work, and despite a different operationalization of neighborhoods from Dautriche et al (2017), we found that wordforms in the real lexica had larger average neighborhood sizes than wordforms in the artificial lexica (see *Figure 25* below). For example, wordforms in the real English lexica averaged 2.56 neighbors, whereas the mean neighborhood sizes in the artificial English lexica ranged from 2.23 to 2.32 (M = 2.28, SD = 0.03). This result is the inverse of our finding regarding homophony—wordforms in the artificial lexica have *more* homophones on average than wordforms in the real lexica. In other words, the artificial lexica appear to optimize for dense concentrations of homophony, while the real lexica appear to optimize for larger neighborhoods. This apparent trade-off can also be illustrated by comparing both the rank-distribution of homophone counts and rank-distribution of neighborhood sizes across the real and artificial lexica (see *Figure 26*).

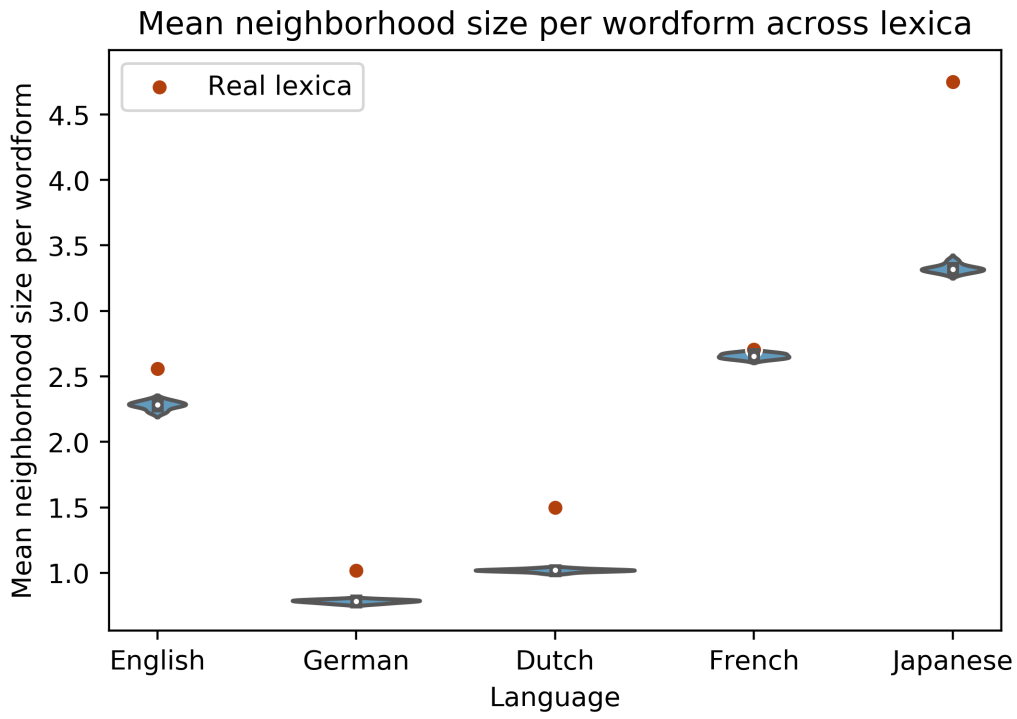


Figure 25: Consistent with previous work (Dautriche et al, 2017), wordforms in the real lexica (shown by orange dots) have larger lexical neighborhoods (i.e., the set of words differing in exactly one phoneme) on average than wordforms in the artificial lexica (shown by violin plots). Note that this is true even in French, where the values are closest: wordforms in the real French lexicon have 2.71 neighbors on average, whereas wordforms in the artificial lexica have approximately 2.66 neighbors on average (SD = .02).

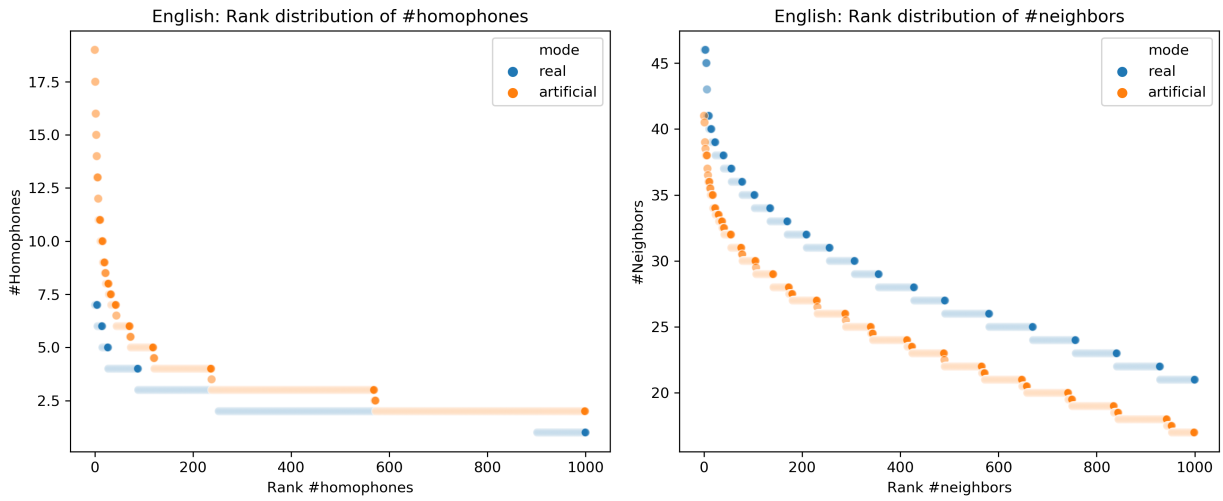


Figure 26: Rank-distribution of homophone counts (left) and rank-distribution of neighborhood sizes (right) across the real and artificial English lexica. The most homophonous wordforms in the artificial lexica are more homophonous than equivalently ranked wordforms in the real lexica. Conversely, the wordforms with the largest neighborhoods in the artificial lexica still have smaller neighborhoods than equivalently ranked wordforms in the real lexica.

Taken together, these findings are broadly consistent with the hypothesis that real lexica could be subject to a pressure against oversaturating the same wordform with too many meanings, and this selection against homophony could instead result in the creation of lexical neighbors. Of course, a similar effect could be achieved not through selection against high levels of homophony but rather from a positive pressure towards large neighborhoods, i.e., a "clumpy" lexicon. As Dautriche et al (2017) argue, dense lexical neighborhoods may have many beneficial consequences, e.g., for word learning (Coady & Aslin, 2003; Storkel et al, 2006; though see also Swingley & Aslin, 2007) and word production (Vitevitch et al, 2002; Vitevitch & Sommers, 2003). It is impossible to know from the current work whether the disparity between the real and artificial lexica is due to a direct pressure in real lexica against oversaturation that results in

dense neighborhoods, or a positive selection for dense neighborhoods that results in less homophony. Future work could explore this potential trade-off at both the psychological level of explanation (e.g., whether learners make errors when learning homophones that lead to the creation of near neighbors), and by simulating such pressures during the lexicon generation process (e.g., whether a direct pressure in favor of large neighbors reduces the number of homophones, or whether a direct pressure against over-saturation increases neighborhood size).

Homophones could conceivably be reduced in real lexica through other, more indirect mechanisms as well. Notably, many human languages have rich morphological structure, allowing them to flexibly combine existing morphemes to construct novel meanings. While the real lexica we analyzed excluded wordforms derived via inflectional morphology, they did not exclude derivational morphology (e.g., adding the suffix *-ify* to the adjective *humid* creates the verb *humidify*; adding the suffix *-ness* to the adjective *happy* creates the noun *happiness*). Morphological compositionality allows speakers to convey new meanings without coining entirely new wordforms—but it also avoids the need to reuse existing wordforms for new, unrelated meanings (i.e., homophony). Thus, compositionality represents an efficient mechanism for recycling existing lexical materials that also avoids outright ambiguity. Clearly, wordforms in the artificial lexica were not constructed via processes of morphological composition. Future work could also explore whether parameterizing these artificial lexica according to the morphology of the underlying real lexicon would decrease the overall homophony, and if so, how. (See *Supplementary Analysis 4* for further exploration of the relationship between derivational morphology and homophony in real lexica.)

In addition to real lexica exhibiting a lower upper-bound on homophony overall, we found that their homophones were less optimally distributed—that is, homophones were much

more concentrated among short, phonotactically likely wordforms in the artificial lexica than in their real counterparts. This result is surprising on its face: why do real lexica apparently prefer (at least relative to the phonotactic baselines) to distribute their homophones across less optimal regions of the lexicon? Even if real lexica select against over-saturation, intuition suggests that the homophones that *are* preserved should be concentrated among short, phonotactically likely wordforms. One possible explanation for this result is that the pressures that ordinarily select against homophony are reduced for longer wordforms—there are at least two accounts as to why this may be the case. The first account is that longer wordforms might be more contextually discriminable than short wordforms and are thus more likely to be preserved in the lexicon. If this is true, the distinct senses of homophonous wordforms should be better disambiguated by contextual cues (e.g., some representation of the linguistic context) for longer wordforms. The second account holds that because longer wordforms are comparatively less common than short wordforms, they require less frequent disambiguation. Even if longer wordforms are no more contextually discriminable than short wordforms, they are encountered less often. If a frequent need to disambiguate is one of the factors that selects against homophony—e.g., because disambiguation may incur processing costs, no matter how marginal—homophones should be relatively *more* likely to be preserved among infrequent wordforms than frequent ones. Note that this does not predict that short wordforms have less homophones overall; past work (Piantadosi et al, 2012) has shown empirically that this is not the case. Rather, the penalty against accruing multiple meanings will be proportionately less for longer, less frequent wordforms. Therefore, less frequent wordforms should experience less of a reduction in their *projected* homophony (relative to their phonotactics) than more frequent ones.

As noted above, the artificial lexica are intended as statistical baselines to determine which theoretical parameters are required to explain homophony, not as models of the many other pressures that real lexica are subject to. Thus, our work does not elucidate the developmental or historical mechanisms by which homophones arise, nor the processes by which they might be selected against or be preserved in a lexicon. There are a number of known sources of homophony in real lexica, including sound change and lexical borrowing (Ogura & Wang, 2006; Ke, 2007). Despite some debate about the extent to which homophony-generating sound changes are avoided (Sampson, 2013; Wedel et al, 2013; Sampson, 2015; Yin & White, 2018), there are many attested examples of phoneme losses and mergers resulting in homophony, such as *knight* and *night* in English, or as a consequence of the many phoneme mergers experienced in Middle Chinese (Ke, 2007; Sampson, 2013; Sampson, 2015). Similarly, lexical borrowing can lead to homophony; for example, the English words *sheik* and *chic* were both borrowed from different languages at different time points (16$^{th}$ century Arabic vs. 19$^{th}$ century French, respectively), and both have an identical phonological form (Ke, 2007). A satisfying explanation of homophony at a mechanistic level should incorporate these generative processes—i.e., the *mutations* by which potential homophones are introduced into a lexicon. Such a model should also predict which potential homophones will be selected against (and what form this selection process takes, i.e., whether it is via the avoidance of homophony-inducing mergers (Wedel et al, 2013; Yin & White, 2018) or something else) and which will be preserved. Homophones should be more likely to survive in a lexicon if their meanings are systematically made sufficiently discriminable by context (Dautriche et al, 2018). A better understanding of this process would also yield insights into which sources of contextual information human speakers and

comprehenders routinely sample and deploy for disambiguation, and therefore influence language change.

We began by asking why a system that appears to be optimized for efficient communication (Gibson et al, 2019) contains apparently inefficient properties such as lexical ambiguity. A series of simulations suggests no evidence for a direct selection pressure in favor of homophones. Rather, the concentration of homophony among short, high-probability wordforms can be explained purely as a function of a language's phonotactics and distribution of word lengths, which perhaps themselves are the result of a pressure for efficiency. In fact, real lexica may even select *against* dense concentrations of homophony. We have suggested one mechanism: they might "smooth out" high-probability phonotactic sequences across lexical neighborhoods instead of concentrating these sequences in a single wordform. The product is lexica that are slightly less optimal in phonotactic terms but may better conform to other requirements of humans who need to use them.

**Acknowledgements**

Chapter 5, in full, is a reprint of the material as it appears in *Cognition*. Trott, Sean; Bergen, Benjamin (2020). The dissertation author was the primary investigator and author of this paper.

# CHAPTER 6: CAN A PRESSURE AGAINST HOMOPHONES EXPLAIN PHONOLOGICAL NEIGHBORHOODS?

Why are human languages structured the way that they are? One approach to finding evolutionary causes for contemporary structure seeks to characterize the various *selection pressures* that could plausibly account for the form and content of languages (Richie, 2016). This approach has produced a growing consensus that human lexica are shaped by a pressure for cognitive and communicative efficiency (Gibson et al., 2019; Levshina & Moran, 2021), both in terms of how they carve up semantic domains (e.g., color) (Regier, Kay, & Khetarpal, 2007; Kemp & Regier, 2012; Zaslavsky, Kemp, Regier, & Tishby, 2018; Kemp, Xu, & Regier, 2018), and in the wordforms they contain (Piantadosi, Tily, & Gibson, 2011; Mahowald, Dautriche, Gibson, & Piantadosi, 2018).

But one feature of language that has to date resisted explanation in these terms is the presence of dense *phonological neighborhoods*. Lexica are *clumpy*: they contain dense pockets of wordforms differing in only one sound (e.g., "cat", "bat", and "mat")—typically called *phonological neighbors*—while leaving vast swaths of phonological space entirely unused (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017). From the perspective of communicative efficiency, this clumpiness may be surprising; allowing wordforms to cluster in particular regions of phonological space——instead of distributing them more evenly——has been found to increase the likelihood of misperceiving one wordform for another, potentially even impairing comprehensibility (Vitevitch & Luce, 1998).[29]

One explanation for the prevalence of neighborhoods comes from *phonotactics*. Each language has certain rules about which sounds can start and end a word, which sounds can occur

[29] Note that contradictory results have been obtained in Russian and Spanish, in which dense neighborhoods may actually facilitate word perception (Vitevitch & Rodr´ıguez, 2005; Arutiunian & Lopukhina, 2020).

in which sequence, and so on (Frisch, Large, & Pisoni, 2000; Bailey & Hahn, 2001; Vitevitch & Aljasser, 2021). Phonotactic rules sharply constrain the space of legal words in a language, simplifying the speaker's task of selecting and producing words. And phonotactics may also account for some of the clumpiness oberved in human languages. However, recent work (Dautriche et al., 2017) has found that phonotactics alone cannot fully account for the neighborhood density of real human lexica: across four languages (English, Dutch, German, and French), phonological neighborhoods are still larger than one would expect in a lexicon whose wordforms were determined purely by the phonotactics of that language (Dautriche et al., 2017). What accounts for this gap?

A natural explanation is that dense phonological neighborhoods are directly *selected for*, i.e., they increase cognitive or communicative efficiency in some way. Indeed, there is some evidence that dense neighborhoods may facilitate both word learning (Storkel, 2004; Storkel, Armbruster, & Hogan,¨ 2006; Coady & Aslin, 2003; Jones & Brandt, 2020; Fourtassi, Bian, & Frank, 2020; Jones & Brandt, 2019) and word production (Vitevitch, 2002; Vitevitch & Sommers, 2003). If this interpretation is correct, it suggests that the possible benefits of dense neighborhoods (facilitation of word learning and production) "outweigh" the challenges they may pose for comprehension (Vitevitch & Luce, 1998). Thus, under this view, neighborhoods are the result of a positive selection pressure—above and beyond the phonotactics of a language.

Another possibility, however, is that dense neighborhoods are the *byproduct* of other properties or selection processes that operate over real human lexica. The fact that neighborhoods appear to confer a benefit on lexical acquisition and production does not entail that they were selected for this function; there are numerous examples in evolutionary biology of apparently adaptive traits that emerged at least partially as a byproduct of other selection

pressures (Gould & Lewontin, 1979). Below, we introduce one such candidate pressure—a selection pressure against homophony—and describe how it could result in lexica with dense phonological neighborhoods, even without a direct selection pressure for clumpiness.

**Real Lexica Select Against Over-Saturation**

There has been a good deal of attention recently on why ostensibly efficient communication systems would evolve to contain homophony, i.e., wordforms with distinct, unrelated meanings (Piantadosi, Tily, & Gibson, 2012). Several papers (Trott & Bergen, 2020; Caplan, Kodner, & Yang, 2020) have adopted the approach of building *phonotactic baselines* (Dautriche et al., 2017) to ask how much homophony one should expect to find purely as a function of a language's phonotactics. That is, if wordforms were randomly sampled (with replacement) from phonotactic space, how frequently would different meanings be assigned to the same wordform?

These phonotactic baselines have been able to account for both the amount and distribution of homophony. But surprisingly, real human lexica actually have *fewer* homophones per wordform than their artificial, phonotactic counterparts (Trott & Bergen, 2020), and this homophony is more evenly distributed across the lexicon, i.e., across longer and more illformed wordforms, than one would expect (Trott & Bergen, 2020; Caplan et al., 2020).

A natural explanation for the gap in homophony is that real lexica are subject to a pressure against saturating the same wordform with too many meanings. A few notes of clarification are required here. First, all spoken languages appear to display homophony, so any hypothesized pressure against homophony must not be categorical (Sampson, 2013). Second, what all languages studied to date share is an apparent resistance to *over-saturation*, i.e., the number of meanings loaded onto the same wordform, relative to what would be expected from a phonotactic baseline. This is despite the fact that some languages (English and Japanese in

particular) have a higher *rate* of homophony (i.e., more wordforms with at least one meaning) than baselines (Trott & Bergen, 2020).[30] Taken together, these facts suggest that real lexica may be subject to a *smoothing* process: rather than concentrating many meanings in the highest-probability wordforms—which could impede communication—real lexica may distribute these meanings more evenly across phonotactic space (Trott & Bergen, 2020), which could result in larger neighborhoods.

Could smoothing create larger neighborhoods? If real lexica prefer wordforms with high phonotactic probability, as they appear to, and if at the same time they also select against over-saturating the same high-probability wordform, then they should be more likely to instead select other high-probability (but not overly homophonous) wordforms in adjacent phonological space. Under this account, the distribution of wordforms across phonological space would be determined by two primary factors:

A pressure to use well-formed phonological sequences, i.e., those with high phonotactic probability.

A pressure against over-saturating the same wordform with an excess of meanings.

Critically, this pair of pressures together could result in larger phonological neighborhoods than either of them would independently, even while not directly selecting for dense neighborhoods. Instead of sampling the same high-probability wordform (e.g., “gap”) many times, this process would sample from similarly high-probability regions of phonotactic space, which—simply because of the previously established connection between phonotactic probability and neighborhood density (Dautriche et al., 2017)—would select wordforms that are

[30] This may also partially account for the mixed results reported in more recent work (Pimentel, Meister, Teufel, & Cotterell, 2021).

more likely than chance to be neighbors of existing words. In the aggregate, this would indirectly produce denser neighborhoods.

This explanation—a pressure against oversaturation of individual wordforms increases neighborhood density—has several things to recommend it a priori. First, the pressure against oversaturation is itself independently motivated, as described above. But second, it could also account for a *dissociation* between homonymy and neighborhood size observed in past work (Dautriche et al., 2017; Trott & Bergen, 2020). Across five languages tested (English, Dutch, German, French, and Japanese) by two groups, real human lexica consistently have larger neighborhoods but fewer homonyms than their phonotactic baselines. Finding a single explanation for both effects is desirable from the perspective of theoretical parsimony; rather than positing multiple, distinct pressures to explain different results—a pressure *against* homophony (Trott & Bergen, 2020) and a pressure *for* denser neighborhoods (Dautriche et al., 2017)—a single pressure could in principle explain two apparently unrelated phenomena, i.e., "filling two needs with one deed".[31]

Under this alternative account, dense neighborhoods may still provide benefits to word learning and production (Storkel, 2004; Storkel et al., 2006; Vitevitch, 2002). However, these advantages would not be causally responsible for larger neighborhoods, but rather, would be a kind of "positive externality" created by a selection pressure against homophones.

**Current Work**

The central goal of the current work was to ask whether a pressure against homophony—coupled with phonotactic constraints—could explain the distribution of neighborhood sizes observed in real lexica. To our knowledge, this account has not been directly tested.

[31] In principle, a pressure for larger neighborhoods may also explain why real lexica have fewer homophones. This issue is explored in the General Discussion.

We followed the approach taken in past work (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020); for each language of interest, we simulated a series of *baselines*, matched for the phonotactics and distribution of word lengths (as defined by number of syllables) of the target lexicon. Unlike past work, however, we also introduced novel constraints for some of these baselines. Specifically, we introduced an Anti-Homophone pressure, which prevented a wordform from acquiring too many meanings and forced the baselines to conform to the rank distribution of homophones found in the real lexicon.[32]

We then compared two measures of neighborhood size (Mean and Maximum Neighborhood Size) across the real lexica and their baselines. Our question was to what extent these constraints—phonotactics, and a pressure against homophony—were *sufficient* to account for neighborhood density in real lexica. Critically, a demonstration of sufficiency would not *disconfirm* the possibility that real lexica are subject to a pro-neighborhood pressure. Rather, it would serve as a proof-of-concept that there are alternative (and possibly more parsimonious) routes that could account for the size of neighborhoods in real lexica.

All materials and code are available on GitHub: https://github.com/seantrott/neighbors_lexica.

**Methods**

**Materials.**

We analyzed five languages: English, Dutch, German, French, and Mandarin. To do this, we relied on lexical resources that contained phonological information for each *lemma* of a lexicon. We used CELEX (Baayen, Piepenbrock, & Gulikers, 1996) for English, Dutch, and

[32] We did not attempt to model the specific cognitive or diachronic mechanisms by which homophony avoidance might come about, e.g., through the inhibition of homophony-producing sound changes (Wedel, Kaplan, & Jackson, 2013; Wedel, Jackson, & Kaplan, 2013); this topic is explored more in the General Discussion.

German; Lexique (New, Pallier, Brysbaert, & Ferrand, 2004) for French; and the Chinese Lexical Database for Mandarin (Sun, Hendrix, Ma, & Baayen, 2018).

To ensure that our analyses were consistent with previous work (Trott & Bergen, 2020; Piantadosi et al., 2012), we restricted our analysis to lemmas. We also removed wordforms containing hyphens, spaces, or apostrophes, as well as proper nouns. The final number of lexical entries (i.e., lemmas) for each real lexicon was: 41887 entries in English, 67583 entries in Dutch, 51718 entries in German, 43782 in French, and 45552 in Mandarin.

**Building Phonotactic Models.**

To model the phonotactic rules of each language, we fit a series of *n*-phone Markov Models to each lexicon (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020). By observing the entire set of wordforms in a language, an *n*phone model can learn statistical contingencies such as which phonemes are most likely to start and end a wordform, and which phonemes are most likely to follow the previous $n - 1$ phonemes.

Following past work (Trott & Bergen, 2020), we identified the optimal *n* for each lexicon using a cross-validation procedure. For each lexicon, we performed a train/test split (75% train, 25% test). Then, we fit a series of *n*-phone models ranging from $n = 1$ to $n = 6$ on the training set, and used these trained models to calculate the phonotactic probability of wordforms in the test set. Importantly, we performed this procedure 10 times for each value of *n*, to ensure that the results were not too sensitive to a particular train/test split. The optimal *n* was defined as the value that maximized the probability of wordforms in the held-out test set—i.e., large enough to capture the appropriate dependencies, but not so large that it overfit to the training set. This procedure resulted in $n = 5$ for English, Dutch, and German; and $n = 4$ for French and Mandarin. (Note that tones were treated as phonemes in the phonotactic model; exploratory analyses

suggest that the *n*phone model captured statistical regularities in which tones co-occurred with the internal structure of the corresponding syllable, but future work could ask about the impact of conditioning tones on particular segments of the preceding syllable (Kirby, 2021).)

Finally, we fit an *n*-phone model to each lexicon using all unique word types (rather than the 75% training set). (Word types, rather than tokens, were chosen to be consistent with past work (Piantadosi et al., 2012; Trott & Bergen, 2020), and to avoid conflating phonotactic probability with word frequency.)

**Phonotactic Baselines.**

Following past work (Dautriche et al., 2017; Trott & Bergen, 2020; Caplan et al., 2020), we used the trained phonotactic models to simulate a series of phonotactic baselines for each language. Unlike past work, we built three different types of baselines (described below), with ten versions for each baseline (for a total of thirty baselines per language).

**Neutral Baselines**. The procedure for generating Neutral baselines was identical to the procedure adopted in past work (Trott & Bergen, 2020). We first identified the number of lemmas (not wordforms) per word length (e.g., the English lexicon has 7,706 monosyllabic lemmas). Then, we used the phonotactic model to generate novel wordforms; each artificial lexicon was constrained to have the same distribution of words per word length as the real lexicon. For example, if an artificial lexicon already had the maximum number of monosyllabic words allowed, future monosyllabic words generated by the model would be discarded. This procedure was continued until the artificial lexicon had the same number as lemmas (not necessarily wordforms) as the real lexicon.

Importantly, there was no constraint on the number of "meanings" a given wordform could acquire (i.e., the same wordform could be sampled an arbitrary number of times, provided more words of that length were required).

**Anti-Homophony Baselines.** The Anti-Homophony Baselines followed an identical procedure as the Neutral Baselines, with one additional constraint: no wordform was allowed to acquire more meanings than the equivalently-ranked wordform in the real lexicon's rank distribution of homophones. That is, if the most homophonous wordform in English had eight meanings, then no wordform in the baseline would be allowed to acquire more than eight meanings— and if the tenth most homophonous wordform had only three meanings, then the tenth most homophonous wordform in the baseline could acquire at most three meanings.

Conceptually, this pressure is akin to "blocking" new meanings from being attached to overly homophonous wordforms, and finding an alternative wordform instead. This is similar (though not identical) to instead adding a new word to the lexicon with some probability $p$, where $p$ decays with the number of meanings already assigned to that wordform.

**Anti-Homophony+ Baselines.** Finally, we considered an alternative implementation of an Anti-Homophony pressure. Rather than simply discarding overly homophonous wordforms, we applied a *sound change* to one of the phonemes in the target wordform.

First, we randomly selected a phoneme in the target wordform to change. Then, we replaced it with a random vowel or consonant (depending on the identity of the phoneme). Finally, to ensure that the resulting wordform was sensible, we evaluated its phonotactic probability; if the wordform's probability was higher than the least-probable wordform in the real lexicon, we added it to the lexicon (provided it also did not have too many homophones).

The motivation for this procedure was that a pressure against homophony may not manifest as "blocking" the offending wordform entirely—overly homophonous wordforms likely have many desirable properties as wordforms of that language (i.e., they are short and well-formed). Thus, this anti-homophony pressure would preserve many of these desirable properties (most of the wordform remains intact) while also avoiding an excess of ambiguity.

Note that this procedure could arguably be interpreted as also implementing an indirect, *pro-neighbor* pressure, given that offending wordforms are directly converted to minimal pairs. However, this pro-neighbor pressure need not necessarily be *pro-neighborhood* per se—if the offending homophones are converted to existing wordforms, the distribution of meanings across wordforms could change without altering the distribution of neighborhood sizes.

**Results**

**Replication of Homophony Results.**

Past work (Trott & Bergen, 2020; Caplan et al., 2020) found that phonotactic baselines without a pressure against homophones exhibited a higher upper-bound of homophony: the Maximum Number of Homophones (i.e., the number of meanings assigned to the most homophonous wordform, minus one) was larger in the baselines than their real counterparts. As depicted in *Figure 27*, we replicated this effect: Neutral baselines consistently contained higher levels of homophony than the real lexica.[33]

[33] The Anti-Homophony and Anti-Homophony+ baselines are excluded from this figure, given that their levels of homophony were constrained not to exceed the real lexicon.

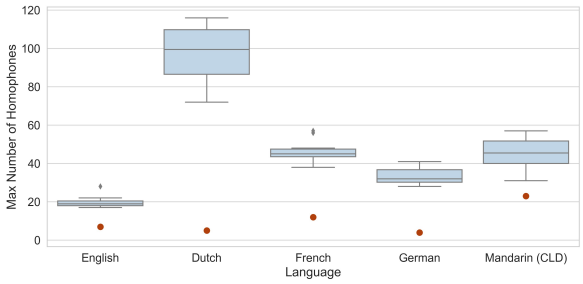


Figure 27: Maximum Number of Homophones across the real lexica and Neutral baselines. Red circles represent the values for the real lexicon.

***Comparing Neighborhood Sizes.***

We used two primary dependent variables to compare the relative density of neighborhoods across real and artificial lexica: Mean Neighborhood Size and Maximum Neighborhood Size.[34] The neighborhood size of a given wordform was defined as the number of wordforms that were exactly one *edit* away, i.e., using either insertion, deletion, or substitution. Thus, the Mean Neighborhood Size was the average phonological neighborhood size across the entire lexicon, while the Maximum Neighborhood Size was the size of the densest neighborhood in a given lexicon.

Consistent with past work (Dautriche et al., 2017; Trott & Bergen, 2020), the real lexica had larger Mean Neighborhood Sizes and Maximum Neighborhood Sizes, compared to the Neutral baselines. For example, the Mean Neighborhood Size in English was 2.51, while the Neutral English baselines ranged from 2.23 to 2.32 ($M = 2.28, SD = 0.03$). Similarly, the Maximum Neighborhood Size in Dutch was 42, while the Neutral Dutch baselines ranged from 25 to 30 ($M = 27.3, SD = 1.89$). This demonstrates that phonotactics alone cannot account for neighborhood density in real lexica.

[34] Equivalent results were obtained using the Total Number of Minimal Pairs within a lexicon, as in past work (Dautriche et al., 2017).

Yet as depicted in *Figure 28*, this gap largely disappeared (or in some cases, reversed) with the introduction of a pressure against over-saturation. Across all languages, the Mean Neighborhood Size was at least as large in the Anti-Homophony baselines. For example, in English, the Mean Neighborhood Size of the Anti-Homophony baselines ranged from 2.52 to 2.59 ($M = 2.54, SD = 0.03$) (recall that the value for the real English lexicon was 2.51). In some languages (e.g., Dutch and German), the Anti-Homophone baselines actually had *larger* neighborhoods on average. The gap was also attenuated for Maximum Neighborhood Size (see *Figure 29*). However, the largest neighborhoods in real lexica tended to be slightly larger than the median value in the baselines (with the exception of French).

Surprisingly, the Anti-Homophony+ baselines exceeded both the Mean and Maximum Neighborhood Sizes of their real counterparts, sometimes to a very large degree (e.g., in French and Dutch). Further, the Anti-Homophony+ baselines *overestimated* the average neighborhood size across all languages tested.

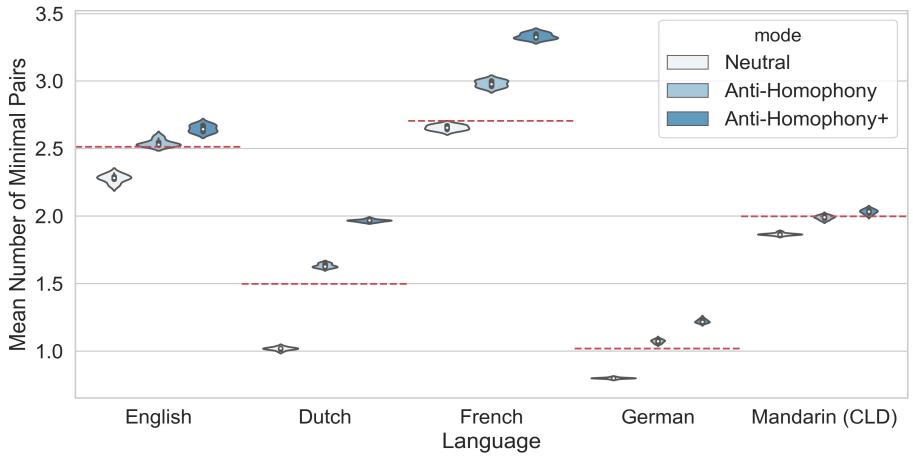


Figure 28: Mean Neighborhood Size as a function of language and lexicon type. Red lines represent the value for each real lexicon.

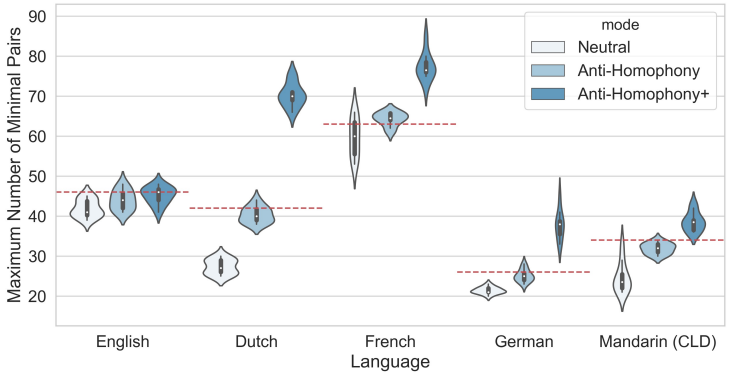


Figure 29: Maximum Neighborhood Size as a function of language and lexicon type. Red lines represent the value for each real lexicon.

In order to quantify which baseline produced the best *fit*, we calculated the Mean Error (ME) between the rank distribution of neighborhood sizes for each real lexicon and its artificial baselines. Mean Error was used (rather than mean absolute or squared error) to reveal the direction of the average error, i.e., whether a given baseline tended to underestimate or overestimate neighborhood sizes on average. As depicted in *Figure 30*, the Neutral baselines generally exhibited the worst fit (with the exception of French), and tended to underestimate neighborhood sizes. The Anti-Homophony baselines produced better predictions, and in fact, had the best fit for every language but French (in which the predicted neighborhood sizes were too large on average). Finally, the Anti-Homophony+ baselines erred on the side of *overestimating* neighborhood sizes, to an even greater degree than the Anti-Homophony baselines.

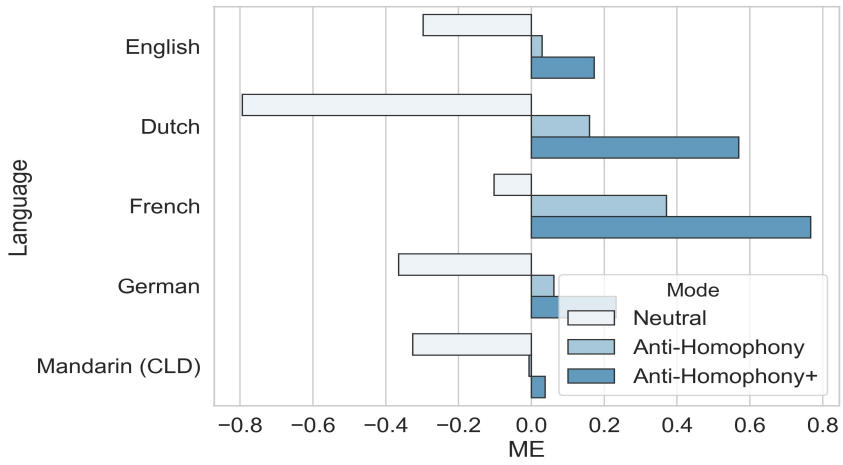


Figure 30: Mean Error (ME) for each baseline. Mean Error was computed by comparing the neighborhood sizes across each real lexicon and its artificial baselines; a score closer to zero corresponds to better fit.

**General Discussion**

We asked whether the distribution of neighborhood sizes in real lexica could be explained by the combination of phonotactic constraints and a pressure against homophony. Past work (Dautriche et al., 2017) found that phonotactics alone were insufficient to account for neighborhood sizes in real lexica, suggesting that real lexica are shaped by a positive selection pressure for larger neighborhoods. This proneighborhood pressure would also be consistent with evidence that dense neighborhoods confer benefits on learning (Coady & Aslin, 2003; Storkel, 2004; Fourtassi et al., 2020) and production (Vitevitch, 2002). The current work replicated this effect, as well as the finding that phonotactics alone tends to *overestimate* the degree of homophony compared to real lexica (Trott & Bergen, 2020).

Critically, however, we found that introducing a pressure against homophony in the baselines resulted in substantially larger neighborhood sizes on average—eroding or even reversing (in French and Dutch) the gap between the real lexica and their baselines (see *Figure 28*). This also resulted in a larger *upper-bound* on neighborhood sizes in the baselines, though

not always larger than the real lexica (see *Figure 29*). Finally, a pressure that converted overly homophonous wordforms to minimal pairs resulted in larger neighborhoods across the board—surpassing the Mean Neighborhood Size of real lexica, and attaining or surpassing the Maximum Neighborhood Size of real lexica.

Thus, a pressure against homophony was in many cases *sufficient* to account for average neighborhood sizes. This means that an explanation for average neighborhood sizes in real lexica need not posit a direct selection pressure for these neighborhoods: the distribution of neighborhood sizes observed in real languages may be the sole result of phonotactics and a pressure against over-saturation. Additionally, the Anti-Homophone+ pressure actually overestimated neighborhood sizes in many cases.

While these results cannot rule out the possibility that neighborhoods are directly selected for (see below), they do demonstrate that a pro-neighborhood pressure may not be a *necessary* part of an explanation. Importantly, this would not be inconsistent with evidence that dense neighborhoods provide benefits to learning and production—but under this account, these benefits would simply be "positive externalities" of a causally unrelated pressure against over-saturation.

**Limitations and Future Work**

The work described here is limited in certain ways. First, the languages tested represent a limited subset of the world's languages. The sample was biased towards Indo-European languages (English, Dutch, German, and French), with one SinoTibetan language (Mandarin), and did not include languages from other major language families such as Austronesian or Niger-Congo. The languages reflect a convenience sample; they are the languages for which we could

obtain lexical resources that contained phonological information at the level of individual lemmas.

A second limitation lies in the measures of neighborhood density used. We used the average and maximum neighborhood size in a lexicon. However, past work (Dautriche et al., 2017) also used more sophisticated measures of the network structure in a lexicon, such as the degree of *transitivity*. Future work in this vein could better quantify how exactly neighborhoods distribute across the lexicon, using tools from network analysis.

Third, as in past work (Dautriche et al., 2017; Caplan et al., 2020; Trott & Bergen, 2020), we used an *n*-phone model to learn the phonotactics of the target language. Recent work has used more sophisticated approaches, such as a generative model (Futrell, Albright, Graff, & O'Donnell, 2017) or LSTM neural network (Pimentel et al., 2021). Future work could ask how adopting an alternative approach to modeling phonotactics changes the distribution of neighborhood sizes in the baselines. That said, recent work (Trott & Bergen, 2022) did find comparable results using an LSTM and *n*phone approach.

Fourth, our approach cannot cannot directly *disconfirm* the theory that real lexica are shaped by a pro-neighborhood pressure. At best, the baselines demonstrate the *sufficiency* of a particular set of constraints in explaining the distribution of neighborhood sizes, absent a direct pro-neighborhood pressure; there may still be *a priori* reasons to prefer a theory that posits such a pressure. The results do suggest that a pressure against homophony can in principle explain two seemingly independent facts—namely, that real lexica have fewer homophones, and larger neighborhoods, than predicted by their phonotactics—but they do not rule out the possibility of alternative explanations.

A fifth, related limitation is that the baselines do not illuminate the causal mechanisms by which an anti-homophony pressure could operate, either at the level of individual communicative constraints or diachronic language change. Future research would benefit from experimental work directly probing these causal mechanisms, e.g., whether errors made during learning homophones (Casenhiser, 2005) could result in minimal pairs. Similarly, researchers could build computational models of how these local pressures interact with changes operating over longer timescales, such as sound change (Wedel, Jackson, & Kaplan, 2013).

Sixth, this work did not consider other important variables, such as *frequency*—both of individual wordforms, and of the distinct lemmas conveyed by those wordforms. This is in part due to limitations of the simulation method used. Employing a different approach, recent work (Trott & Bergen, 2022) discovered several relevant findings: homophony resistance is positively correlated with the frequency of particular *wordforms*, though not necessarily with the relative frequency of their meanings; and further, homophony resistance is highest among wordforms with high neighborhood density—consistent with the results presented here.

Finally, these analyses made two simplifying assumptions. First, meanings were implicitly assumed to be discrete units, with no relation between them. However, meanings are likely at least partially continuous (Elman, 2009; Trott & Bergen, 2021; Li & Joanisse, 2021); further, some meanings are more related (as in polysemy) than others (as in homonymy). Second, forms were assumed to be arbitrarily related to meanings—however, there is considerable evidence (Blasi, Wichmann, Hammarstrom, Stadler, & Christiansen,¨ 2016; Gutierrez, Levy, & Bergen, 2016) that form-meaning´ relationships may be less arbitrary than previously thought. Future work could integrate both lines of thought by using a continuous

representation of the meaning space, and exploring different ways of assigning form-meaning pairings in either systematic or arbitrary ways.

**Conclusion**

Why do real lexica have such large phonological neighborhoods? One explanation is that real lexica are subject to a selection pressure for dense neighborhoods, possibly because dense neighborhoods facilitate word learning (Storkel, 2004; Coady & Aslin, 2003) and production (Vitevitch, 2002; Vitevitch & Sommers, 2003). We pursued another possibility— that dense neighborhoods emerge from the interaction of other constraints operating over real lexica, namely phonotactics and a pressure against individual wordforms acquiring too many meanings (Trott & Bergen, 2020). We tested the sufficiency of this latter account using simulated baselines. Crucially, the combination of phonotactic constraints and an anti-homophony pressure was *sufficient* to account for average neighborhood sizes in real human lexica—demonstrating that a direct selection pressure for neighborhood density is not a *necessary* part of an explanation.

**Acknowledgments**

Chapter 6, in full, is a reprint of the material as it will appear in the *Proceedings of the 44th Annual Conference in Cognitive Science*. Trott, Sean; Bergen, Benjamin (2022). The dissertation author was the primary investigator and author of this paper.

# CHAPTER 7: LANGUAGES ARE EFFICIENT, BUT FOR WHOM?

Languages adapt to the needs of the people who use them. In particular, there is increasing evidence that human languages have evolved in part to facilitate *efficient communication* (Piantadosi et al, 2009; Gibson et al, 2019; Mahowald et al, 2020; Zaslavsky et al, 2018; Regier et al, 2016; Kemp et al, 2018). Pressure for efficiency has been used to explain various features of language, like how they carve up semantic domains among words (Kemp & Regier, 2012; Gibson et al, 2017; Zaslavsky et al, 2018; Conway et al, 2020), as well which wordforms a lexicon contains (Piantadosi et al, 2011; Meylan & Griffiths, 2017; Mahowald et al, 2018). But efficiency involves trade-offs: features that make a language more efficient for *speakers* sometimes make it less efficient for *comprehenders*, and vice versa (Zipf, 1949). How do languages balance the interests of speakers and comprehenders when those interests are misaligned?

In the case of a language's grammatical rules, there is an emerging consensus that languages reflect a trade-off between reducing complexity (i.e., minimizing difficulties in production) and reducing ambiguity (i.e., minimizing difficulties in comprehension). The need to balance these pressures may explain cross-linguistic patterns in word order (Hahn et al, 2020), person marking (Zaslavsky et al, 2021), case marking (Mollica et al, 2020), and more. Moreover, some theories argue that efficiency is best achieved by prioritizing the needs of speakers specifically (Levinson, 2000; MacDonald, 2013). Planning and producing utterances is cognitively expensive: speakers must ultimately translate the concepts they wish to convey into a series of complex motor commands, a process that involves selecting the correct lexical items and arranging them in an appropriate syntactic configuration (Ferreira, 2008; MacDonald, 2013).

The architecture of the language production system is largely tuned towards reducing speaker effort (Ferreira, 2008), and as a consequence, the form of human languages themselves may also be oriented towards *producibility*, rather than *comprehensibility* (Levinson, 2000; MacDonald, 2013). On this view, comprehension is nevertheless possible because comprehenders have a sufficiently less taxing task than speakers (MacDonald, 2013, MacDonald, 2015), and rely on pragmatic inference to decipher under-specified or ambiguous utterances (Levinson, 2000). Of course, some grammatical features may also reflect a pressure for efficient comprehension, such as grammatical gender (Wasow et al, 2013; Dye et al, 2017; Dye et al, 2018). Similarly, the mere fact of grammatical regularity in the first place likely makes communication more robust to noise, which helps with both comprehension and production (Gibson et al, 2013).

There is substantively less consensus when it comes to the lexicon. Although many researchers agree that human lexica are shaped for efficient communication (Piantadosi et al, 2009; Mahowald et al, 2018; Mahowald et al, 2020), it remains unclear whether they favor a pressure for efficient production or efficient comprehension, or whether they are shaped equally by both pressures. The paradigm example of these pressures in conflict is Zipf's *meaning-frequency* law (Zipf, 1945; Piantadosi et al, 2012), the empirical observation that more frequent words are more ambiguous.

On the one hand, this distribution could be interpreted as serving the speaker's needs. It is easier to produce frequent words than infrequent ones (Oldfield & Wingfield, 1965; Dell, 1990), so a lexicon that concentrates meanings among its most frequent wordforms would be more efficient for speakers than a lexicon that distributes its meanings more evenly across wordforms (Zipf, 1949; Piantadosi et al, 2012). Under this view, the meaning-frequency law reflects a pressure for efficient production, which Zipf (1945) termed *unification*. Taken to the extreme,

this pressure—sometimes called *compressibility* (Kirby et al, 2015)—leads to a *degenerate* lexicon, "in which every meaning is associated with a single, shared, maximally ambiguous signal" (Kirby et al, 2015, pg. 88). A maximally degenerate lexicon is often taken as a speaker's ideal because it requires a speaker to remember and produce only a single word, and thus imposes minimal costs on speakers, i.e., it is minimally complex (Zipf, 1949; Zaslavsky et al, 2018).

On the other hand, real lexica are far from maximally degenerate. This is because lexica are also subject to a countervailing pressure, alternatively termed *diversification* (Zipf, 1945), *expressivity* (Kirby et al, 2015), or *informativity* (Zaslavsky et al, 2018), to reduce the burden of comprehension (Zipf, 1945; Wasow, 2013) and ensure clarity of communication (Piantadosi et al, 2012). An incomprehensible language is not particularly efficient—suggesting that the cost of disambiguation should in principle also shape the development of communicative systems. Oversaturating frequent wordforms with many meanings likely incurs costs for comprehenders: even if disambiguation is less costly than production (Levinson, 2000), it does appear to impose at least a marginal increase in processing difficulty (Rayner & Duffy, 1986; Rayner & Frazier, 1989; Blott et al, 2020). And if the most frequent wordforms are also the most ambiguous, then comprehenders will be required to disambiguate more often. But while real lexica do exhibit a relationship between frequency and ambiguity, their most frequent wordforms are not maximally ambiguous, i.e., this relationship is weaker than would be expected by a purely speaker-centric lexicon. Thus, under this view, the empirical relationship between ambiguity and frequency also reflects a pressure to reduce the burden on comprehenders.

Zipf's interpretation is that the empirical distribution of meanings across wordforms represents a *compromise* between these purported pressures (Zipf, 1945; Zipf, 1949). Yet

identifying an equilibrium is only part of an explanation, as it leaves the relative magnitudes of the countervailing pressures indeterminate. It is possible that the pressures are equal in size, as Zipf (1945) suggests. But it could be that a speaker-oriented pressure has ultimately won out—that the equilibrium point is closer to the speaker's ideal than the comprehender's. This view of a Speaker-Oriented Pressure is similar to claims that grammar shows an equivalent bias (MacDonald, 2013). Alternatively, the lexicon may be driven primarily by a Comprehender-Oriented Pressure, biased towards reducing the cost of disambiguation.

Unfortunately, we cannot adjudicate between these competing accounts using the empirical distribution of word meanings alone. In the absence of a suitable baseline, it is impossible to determine whether Zipf's meaning-frequency law is attributable to a bias towards production or a bias towards comprehension, or even whether it can be explained without either such pressure (Caplan et al, 2020; Trott & Bergen, 2020). To date, the observed relationship between wordform frequency and ambiguity has only been compared with a baseline in which there is *no* relationship between wordform frequency and ambiguity (Zipf, 1949; Piantadosi et al, 2012). But such a baseline is indistinguishable from one version of a purely Comprehender-Oriented Pressure, in which meanings are distributed evenly across wordforms, and is thus inappropriate for adjudicating between Speaker-Oriented and Comprehender-Oriented Pressures. Instead, a baseline is required that establishes how many meanings those same wordforms should be expected to accrue just on the basis of other known factors. Previous work has established that even controlling for frequency, shorter and more phonotactically probable words have more meanings (Piantadosi et al, 2012). Using a baseline that incorporates these effects, we can then ask whether the positive empirical relationship between wordform frequency and ambiguity is

larger (reflecting a Speaker-Oriented pressure) or smaller (reflecting a Comprehender-Oriented pressure) than what would be expected without either such pressure.

Here, a conceptual parallel can be drawn to work in evolutionary biology; many traits that appear adaptive for a particular function may have emerged from other, more indirect selective pressures, or even genetic drift (Gould & Lewontin, 1979). This has led to the use of so-called "neutral" models (Alonso et al, 2006) to establish baselines of what to expect in the absence of selection pressures. More recently, neutral models have been applied to cultural evolution as well, to understand which aspects of language change are due to explicit selection and which are better explained by stochastic drift (Newberry et al, 2017). There is some controversy around the question of whether neutral models can be used to provide positive evidence of a causal mechanism (Leroi et al, 2020; Bentley et al, 2012); however, there is general agreement that they are useful for establishing a "null" baseline, against which alternative theoretical models can be compared (Leroi et al, 2020).

Consonant with this line of reasoning, recent work has shown that when the observed distribution of homophony is compared with an appropriate baseline, other apparently efficient distributions of meanings show up in lexica without any explicit pressure for efficiency (Trott & Bergen, 2020; Caplan et al, 2020). Indeed, Trott & Bergen (2020) find that when compared against a suitable baseline that incorporates a lexicon's phonotactics and distribution of word lengths, real human lexica actually have *fewer* homophones than one would expect. Strikingly, this result is consistent with a Comprehender-Oriented Pressure, i.e., one in which homophones are avoided during the course of language change (Wedel et al, 2013a; Wedel et al, 2013b). Importantly, however, because this work used a simulated baseline (i.e., not using real words in the lexicon), it was unable to investigate whether the frequency of actual wordforms in a lexicon

shaped a pressure for or against homophony. This leaves a gap in the literature: could a Comprehender-Oriented Pressure explain Zipf's meaning-frequency law as well?

The logic of our approach below is as follows. First, we establish a suitable baseline that characterizes the expected relationship between wordform frequency and ambiguity in the absence of either a direct production-oriented pressure or a comprehension-oriented pressure. The distribution obtained in this baseline is then compared to the attested distribution in real lexica. If the relationship between frequency and homophony is stronger in real lexica than in the baseline, it is consistent with production-oriented pressures shaping the language; in contrast, a weaker relationship in real lexica is consistent with the language being shaped by a comprehension-oriented pressure. Finally, if the real relationship between frequency and homophony is indistinguishable from the baseline, it suggests either that both pressures are equal in magnitude, or that neither pressure is required to explain how many meanings words of different frequencies have.

This hinges on first establishing a procedure for distributing meanings that is *neutral* with respect to whether it privileges a speaker-oriented pressure to accumulate meanings among frequent wordforms, or a comprehender-oriented pressure to reduce ambiguity among those wordforms. That is, given $M$ meanings and $W$ wordforms, how ought those meanings to be distributed across wordforms in a neutral manner? One candidate for such a neutral procedure is to assign meanings to wordforms according to their phonotactic probability. Although all wordforms of a language must obey the phonotactic rules of that language—i.e., which sounds can begin and end a word, which sounds can occur in which sequence, and so on—some phonological sequences are nonetheless more common across wordforms than others. Wordforms containing very common phonological sequences are considered to have a higher

*phonotactic probability* (Vitevitch & Aljasser, 2021). Critically, phonotactic probability appears to facilitate word production (Vitevitch et al, 2004; Goldrick & Larson, 2008), word recognition and processing (Vitevitch & Luce, 1999; Vitevitch et al, 1999), and word learning (Juszcyk et al, 1994; Munson, 2001; Coady & Aslin, 2004; Storkel, 2002). To our knowledge, there is no evidence that phonotactic probability disproportionately benefits speakers over listeners, or vice versa. Thus, it is reasonable to expect that both speakers and listeners would prefer a lexicon that privileged phonotactically probable wordforms, as opposed to phonotactically improbable ones. (Of course, according to Zipf (1949), speakers might prefer that every meaning is conveyed by a single, high-probability wordform—while listeners might prefer no ambiguity at all. However, the goal of this baseline is not to implement the ideal speaker-oriented or listener-lexicon—it is to construct a lexicon according to neutral principles.)

A second, related reason to distribute meanings according to the phonotactic probability of wordforms is that in real lexica, homophones are disproportionately concentrated among phonotactically probable wordforms (Piantadosi et al, 2012; Trott & Bergen, 2020). This lends further plausibility to the approach being taken: empirically, meanings are attracted to high-probability regions of phonotactic space.

Finally, phonotactic probability correlates with frequency across a number of languages (Bentz et al, 2016; Mahowald et al, 2018; Meylan & Griffiths, 2017). While this is not itself a reason to adopt this baseline, it does tell us *a priori* that even in the absence of a frequency bias, a preferential distribution of meanings according to phonotactic probability would produce a positive correlation between frequency and ambiguity. Importantly, this baseline correlation with frequency would be epiphenomenal in the sense that it emerged from other principles of lexicon design. The central question of the current work is whether the correlation between frequency

and ambiguity in the baseline is *weaker* than the one observed in real lexica (implying a speaker-oriented pressure), or *stronger* than the one observed in real lexica (implying a comprehender-oriented pressure).

## Current Work

Using a neutral baseline, we calculated the Homophony Delta for each wordform in the real lexicon: the difference between how many homophones a wordform *actually* has, and how many homophones it would be *expected* to have, assuming that meanings distributed purely according to phonotactic probability. We then asked whether the relationship between Homophony Delta and Frequency was positive (as predicted by a speaker-centric pressure) or negative (as predicted by a comprehender-centric pressure).

To calculate the expected number of homophones, we first calculated the phonotactic probability of each wordform using an n-phone model[35]. We then multiplied each wordform's phonotactic probability by the number of meanings for words of that length (see the Methods section below for more details on how the number of meanings was calculated). This ensured that the distribution of meanings across word lengths was matched across each of the real lexica and their neutral baselines; for example, if the real English lexicon has 7,706 meanings distributed among its monosyllabic wordforms, the English baseline would do the same. Finally, we subtracted a wordform's expected number of homophones from the number of homophones a wordform actually has. A positive value of Homophony Delta indicates that a wordform has *more* homophones than expected, and a negative value indicates that it has *fewer*. We repeated this process across six target languages: English, Dutch, German, French, Japanese, and Mandarin.

[35] See *Supplementary Analysis 5* for a replication of the primary results using a measure of phonotactic probability calculated using an LSTM.

The accounts outlined above make opposing predictions about the relationship between Frequency and Homophony Delta. Given that more frequent wordforms are easier and faster to produce (Oldfield & Wingfield, 1965; Dell, 1990), a pressure to minimize speaker effort should result in frequent wordforms acquiring more meanings than their phonotactics would predict. Thus, Frequency should exhibit a *positive* relationship with Homophony Delta. On the other hand, concentrating meanings in the most frequent wordforms results in a language requiring more frequent disambiguation by comprehenders. Such a lexicon would impose a larger average disambiguation cost than one that distributed its meanings more evenly across wordforms. Thus, a pressure to minimize comprehender effort predicts a *negative* relationship between Frequency and Homophony Delta. Finally, it is possible that these pressures are roughly equal in size, or even that neither pressure plays a role at all—i.e., that phonotactic plausibility and length is the sole determinant of homophony. In both cases, the relationship between Frequency and Homophony Delta should be statistically indistinguishable from zero.

All data and code necessary to reproduce the analyses described here can be found on GitHub: https://github.com/seantrott/homophony_delta.

**Methods**

**Materials**. We analyzed lexica from six languages: English, Dutch, German, French, Japanese, and Mandarin Chinese. Importantly, we restricted our analysis to the unique *lemmas* of each language. This means that inflectional variants (e.g., "dogs") would not be included as distinct entries, whereas distinct meanings of the same wordform (e.g., *water.n* and *water.v*) would be listed separately, with separate frequency estimates for each lemma. For determining which meanings counted as distinct lemmas, as well as the frequencies of those lemmas, we relied on lexical resources for each language.

For English, Dutch, and German, we used the CELEX lexical database (Baayen et al, 1995). For French, we used the French Lexique (New et al, 2004). For Japanese, we used the Japanese CallHome Lexicon (Kobayashi et al, 1996). For Mandarin Chinese, we used the Chinese Lexical Database (Sun et al, 2018); we also conducted the same analysis (and obtained qualitatively identical results) using the Mandarin CallHome Lexicon (Huang et al, 1996), which is included in the Supplementary Materials. We removed wordforms containing hyphens, spaces, or apostrophes, as well as proper nouns (in the case of the Mandarin Chinese lexica). The number of unique wordforms (i.e., after collapsing across distinct entries) in each lexicon was as follows: 35,107 English wordforms, 50,435 German wordforms, 65,260 Dutch wordforms, 37,278 French wordforms, 40,449 Japanese wordforms, and 41,009 Mandarin Chinese wordforms (with 45,871 in the Mandarin CallHome lexicon).

Frequency estimates for English, Dutch, and German were taken from CELEX; respectively, these frequency estimates were in turn based on the COBUILD (approximately 18 million words), INL (approximately 40 million words), and Mannheim (approximately 5 million words) corpora (Sinclair, 1987; Kruyt & Dutilh, 1997; Kupietz & Keibel, 2009). Note that we also replicated the analyses described here using the SUBLTEX estimates of word frequency, and obtained qualitatively identical results (i.e., a negative relationship between Log Frequency and Homophony Delta; see *Supplementary Analysis 4*) for a description of those results. The lexica for French and Mandarin Chinese already contained by-lemma frequency estimates. The corpus sizes from which these estimates were obtained were, approximately: 14.8M (for French) and 120M (for Mandarin). The frequency estimates for Japanese wordforms were taken from the Japanese CallHome Lexicon, with a total of approximately 690K tokens. In each language, if by-lemma frequency estimates were available for a given wordform, we summed these estimates to

calculate the total frequency of that wordform. Note that the Japanese lexicon did not contain reliable by-lemma frequency measures—thus, for Japanese, we used the *mean* frequency for each lemma corresponding to a given wordform. However, the results reported below are qualitatively identical using the sum of lemma frequencies. Additionally, because we would eventually calculate the log of each frequency, we incremented each frequency value by 1, to ensure that no wordforms had a frequency of 0. Additionally, for the French lexicon specifically, frequency values were multiplied by 14.8 (given that Lexique normalized the book frequency estimates to 14.8).

Finally, the frequency estimates reflect a mixture of spoken and written text, depending on the language. The English COBUILD corpus consists primarily of written language (approximately 5% is spoken), as do the Dutch INL (approximately 9% is spoken) and German Mannheim (0% is spoken). The Chinese Lexical Database frequency estimates combine two written sources: the Leiden Weibo Corpus (Van Esch, 2012) and the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). For French, we relied on frequency estimates from a corpus of written books (New et al, 2004). Finally, frequency estimates for the Japanese CallHome Lexicon are based solely on spontaneous spoken speech (Kobayashi et al, 1996).

**Calculating Phonotactic Probability**. For each lexicon, we built an *n*-phone Markov Model that approximated the phonotactics of the target language. We adapted the code and procedure used in previous work (Dautriche et al, 2017; Trott & Bergen, 2020).

Given some value of *n* (e.g., 2), an *n*-phone model can use the set of wordforms[36] in a lexicon to learn which phoneme characters occur in which positions and in which sequence; for example, in English, such a model would learn that the sequence *bn–* never occurs at the start of a wordform. Such a model can then be used to compute the probability of an entire wordform, which is defined as the product of all the transitional probabilities between each phoneme in that wordform (including the START and END symbols). We identified the appropriate chain length (i.e., value of *n*) for each language using a cross-validation procedure—the optimal *n* was defined as the model that, when trained on a set of real wordforms (e.g., 75% of a lexicon), maximizes the probability of held-out wordforms (e.g., the remaining 25%). This cross-validation procedure was identical to the one described in Trott & Bergen (2020), and determined the optimal models to be 5-phone models for English, Dutch, and German, and 4-phone models for Japanese, French, and Mandarin Chinese.

We then calculated the phonotactic probability of each wordform in each lexicon using 1000-fold cross-validation. We divided each lexicon into 1000 "folds" (each containing roughly 0.1% of the entire set of wordforms). Then, for each fold, we trained an *n*-phone model on the remaining 99.9% of the lexicon, and evaluated the phonotactic probability of the wordforms in the target fold. This allowed us to produce estimates of phonotactic plausibility from a model that never directly observed the wordforms in question—only other wordforms resembling them to varying degrees. As in past work (Dautriche et al, 2017; Trott & Bergen, 2020), we also

[36] Note that these models were trained using the set of unique *types* (individual wordforms), rather than *tokens* (actual instances of each wordform in a text corpus), to avoid conflating phonotactic probability with frequency.

assigned non-zero probability to unobserved phoneme sequences using Laplace smoothing with the parameter set to .01.

Finally, we used these probabilities to calculate the phonotactic surprisal of each wordform, which is defined as the negative log probability (note that we used $log_{10}$)—i.e., less probable phonotactic sequences will have higher phonotactic surprisal. Because phonotactic surprisal is correlated with length, we divided surprisal by the number of phonemes in the wordform to obtain a Normalized Phonotactic Surprisal measure, as in Piantadosi et al (2012).

Note that recent work (Pimentel et al, 2021) has found that an LSTM provides a better measures of phonotactic probability, and is less prone to overfitting, than an n-gram model. We have replicated the primary results described below using an LSTM with qualitatively identical results; see *Supplementary Analysis 5* for more details.

**Calculating Actual Number of Homophones**. Following past work (Piantadosi et al, 2012; Trott & Bergen, 2020), we calculated the Actual Number of Homophones for a given wordform, $A(w_i)$, by identifying the number of distinct lexical entries with the same phonological form, then subtracting one. Note that this measure would include both homographic (e.g., "baseball *bat*" vs. "furry *bat*") and heterographic (e.g., "juicy *steak*" vs. "wooden *stake*") homophones. In the latter case, the wordform */steɪk/* has three entries, so the Actual Number of Homophones is two.

**Estimating Expected Number of Homophones**. To estimate a wordform's Expected Number of Homophones, we calculated the number of meanings each wordform should be assigned if meanings were assigned purely on the basis of phonotactic plausibility alone. We also sought to control for word length, so the procedure described below was performed separately for words of varying lengths (e.g., 1-syllable, 2-syllable, etc.).

First, we normalized a wordform's phonotactic probability, $p_i$, to the number of meanings, *M,* distributed among wordforms of that length. To do this, we calculated the sum of those wordforms' probabilities—typically much less than 1, depending on the smoothing parameter and number of wordforms in question—then divided each probability $p_i$ by that sum. This produced a set of normalized wordform probabilities such that they summed to 1, which ensured that the sum of expected number of meanings (*M'*) distributed among some set of wordforms would equal the actual number of meanings (*M*). After this normalization procedure, the monosyllabic wordform */steɪk/* ends up with a normalized probability of 0.0009.

Then, for each wordform, we multiplied its normalized probability by *M*, the number of meanings available for wordforms of that length. This yielded the expected number of meanings. For example, the normalized probability for the wordform */steɪk/* (0.0009) would be multiplied by the number of meanings available for monosyllabic wordforms (7,706), yielding the expected number of meanings (approximately 6.94).

Finally, to calculate the Expected Number of Homophones, we simply subtracted one from the expected number of meanings (as in the real lexicon); the wordform */steɪk/* would thus have approximately 5.94 homophones. This is illustrated in the equation below, where $w_i$ refers to a given wordform, $M_i$ refers to the number of meanings expressed by wordforms of that length, $p_i$ refers to the normalized probability of that wordform and $E(w_i)$ refers to the Expected Number of Homophones.

$$\mathrm{E}(w_i) = M_i * p_i - 1$$

Note that unlike in the real lexicon, this procedure can yield non-integer values for a wordform's Expected Number of Homophones; occasionally these values are even negative, if the expected number of meanings is below 1. We chose not to "correct" these values (i.e., round

them to the nearest integer), because doing so would no longer ensure equivalence between the *actual* and *expected* number of meanings distributed among wordforms of a certain length. Because our primary interest is in the relative differences between expected and actual numbers of meanings, the absolute value of Expected Number of Homophones should not impact the interpretation of results. (See *Supplementary Analysis 6* for an alternative approach ensuring that wordforms are assigned an integer number of meanings.)

**Calculating Homophony Delta**. Homophony Delta, i.e., $HD(w_i)$, was defined as the difference between a wordform's Actual Number of Homophones, i.e., $A(w_i)$, and that wordform's Expected Number of Homophones, i.e., $E(w_i)$:

$$HD(w_i) = A(w_i) - E(w_i)$$

We subtracted the latter estimate (described above) from the former, obtained from the real lexica. Thus, a negative value means that wordform has *fewer* homophones than predicted by its phonotactics, while a positive value means that a wordform has *more* homophones than predicted by its phonotactics. For the wordform */steɪk/*, the Actual Number of Homophones is 2, while the Expected Number of Homophones is 5.94, so the Homophony Delta would be -3.94. Put another way: the wordform */steɪk/* has approximately 3.94 fewer homophones than predicted by its phonotactics.

**Results**

**Homophony and Frequency.**

For each language, we constructed a linear regression model with Homophony Delta as the dependent variable, and Log Frequency, Number of Syllables, and Normalized Phonotactic

Surprisal[37] as predictors. We were primarily interested in the effect of Log Frequency, which we focus on below; given that frequency is correlated with word length and phonotactic probability, we included Number of Syllables and Normalized Phonotactic Surprisal as covariates to identify and isolate the variance explained by Frequency specifically[38]. All analyses were performed in R version 3.6.3 (R Core Team, 2020).

Log Frequency exhibited a significant, negative relationship with Homophony Delta across all six languages: English [$\beta = -0.49$, $SE = 0.07$, $p < .001$], Dutch [$\beta = -1.85$, $SE = 0.07$, $p < .001$], German [$\beta = -1.28$, $SE = 0.1$, $p < .001$], French [$\beta = -0.45$, $SE = 0.05$, $p < .001$], Japanese [$\beta = -1.73$, $SE = 0.11$, $p < .001$], and Mandarin Chinese [$\beta = -0.28$, $SE = 0.02$, $p < .001$]. The magnitude of this relationship, and the absolute values of Homophony Delta, varied considerably across languages; for example, the most frequent wordforms in Dutch have much larger negative values of Homophony Delta than the most frequent wordforms in French or Japanese. Crucially, however, the overall relationship was negative in each of the languages we considered: frequent wordforms consistently have fewer homophones than predicted by their phonotactics. Because we modeled frequency as logarithmic, these coefficients can be interpreted as representing the expected reduction in homophony (relative to a wordform's phonotactics), given each order of magnitude increase in frequency. For example, in English, the coefficient estimate for Frequency is *-0.49*; this means that an increase in frequency from 10 to

[37] Because Number of Syllables is correlated with Phonotactic Surprisal, we followed the procedure described in Piantadosi et al (2012) and divided Phonotactic Surprisal by the number of phonemes in a wordform, which we called Normalized Phonotactic Surprisal.

[38] Note that Frequency is correlated with Number of Syllables, and the presence of collinearity between predictors can sometimes lead to suppression or enhancement of parameter estimates (Wurm & Fisicaro, 2014). To check for collinearity, we calculated the variance inflation factor (VIF) for the complete model for each language, and found that all VIF scores were below 1.5, which suggests that collinearity is not necessarily a concern in this case.

100 would predict a *0.49* decrease in how many homophones a given wordform has, relative to its phonotactics.

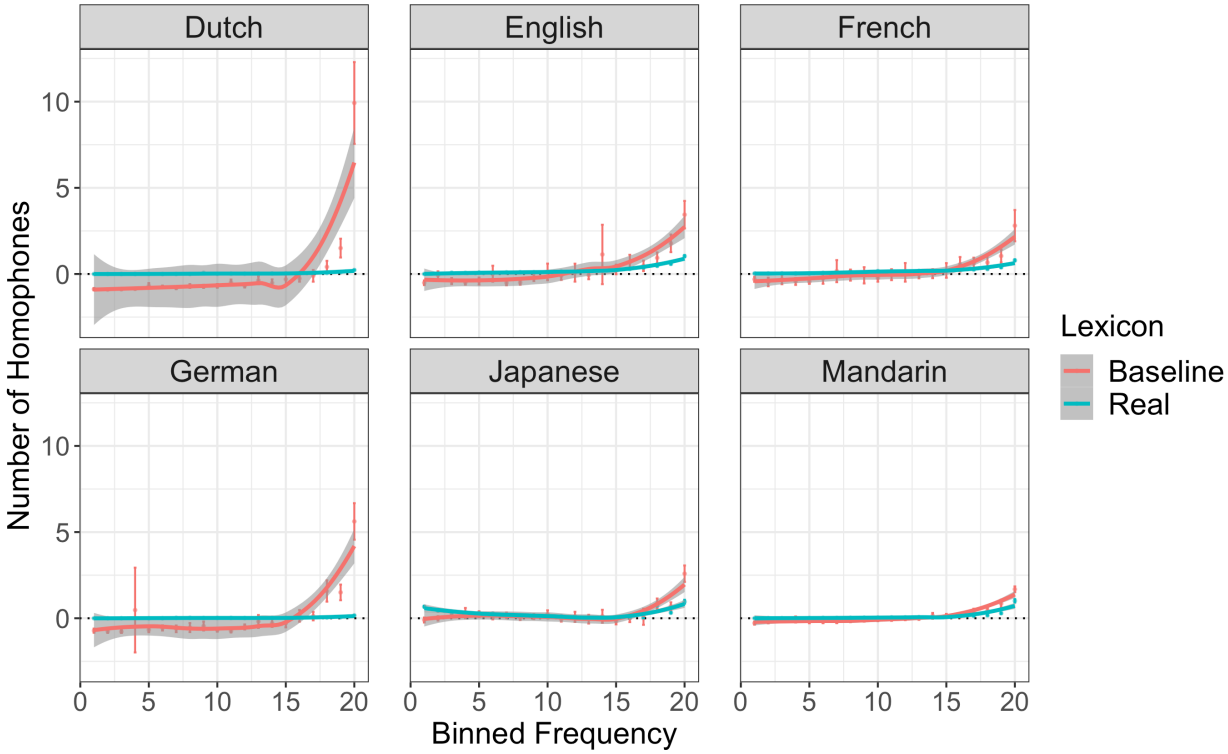


Figure 31: Across all six languages, the most frequent wordforms have fewer homophones in actuality (Real) than predicted by their phonotactics (Baseline). Higher values of Binned Frequency correspond to more frequent words. Error bars are one standard error.

This is best illustrated by *Figure 31*, which directly compares the actual and expected number of homophones for each of 20 frequency bins. Across all languages, frequent wordforms have fewer homophones in actuality than expected. That is, although each language exhibits the well-attested, positive relationship between wordform frequency and ambiguity[39]—i.e., Zipf's

[39] See *Supplementary Analysis 3* for an analysis illustrating that Zipf's meaning-frequency law replicates across all six languages.

*meaning-frequency law* (Zipf, 1945)—this relationship is considerably weaker than one would expect if meanings were assigned purely on the basis of phonotactic probability and length.

In addition to the negative relationship between Log Frequency and Homophony Delta, Normalized Phonotactic Surprisal exhibited a significant, positive correlation with Homophony Delta across all six languages: English *[$\beta = 3.58$, $SE = 0.13$, $p < .001$]*, Dutch *[$\beta = 4.04$, $SE = 0.17$, $p < .001$]*, German *[$\beta = 3.42$, $SE = 0.18$, $p < .001$]* French [$\beta = 3.18$, $SE = 0.11$, $p < .001$], Japanese [$\beta = 2.92$, $SE = 0.09$, $p < .001$], and Mandarin Chinese [$\beta = 2.71$, $SE = 0.07$, $p < .001$]. The most phonotactically plausible wordforms in real lexica have *fewer* homophones than predicted by their phonotactics alone. This is not surprising, given that our baselines assumed that phonotactic probability was the sole determinant of homophony—if the distribution of homophones in real lexica is influenced by any other factors, then the resulting relationship should be weaker than in our baselines.

More surprising is the observation that Number of Syllables was positively correlated with Homophony Delta across five of the six languages (all but Mandarin Chinese): English *[$\beta = 0.73$, $SE = 0.07$, $p < .001$]*, Dutch [$\beta = 0.4$, $SE = .07$, $p < .001$], German [$\beta = 0.36$, $SE = 0.07$, $p < .001$], French [$\beta = 0.56$, $SE = 0.05$, $p < .001$], and Japanese [$\beta = 0.13$, $SE = 0.02$, $p < .001$]. In other words, short wordforms in these languages were less ambiguous than expected, given their phonotactics. The coefficient in Mandarin was not significant after correcting for multiple comparisons ($p > .1$). Across all languages, however, short wordforms were no *more* homophonous than one would expect (i.e., no language had a negative coefficient for Number of Syllables); this finding is consistent with past work (Trott & Bergen, 2020; Caplan et al, 2020) suggesting that the empirical relationship between length and homophony is not necessarily a product of a speaker-centric pressure to reuse short wordforms—indeed, in some languages,

short wordforms have fewer homophones than one would otherwise expect. See *Figure 32* for the complete distribution of parameter estimates (and standard errors) across lexica.

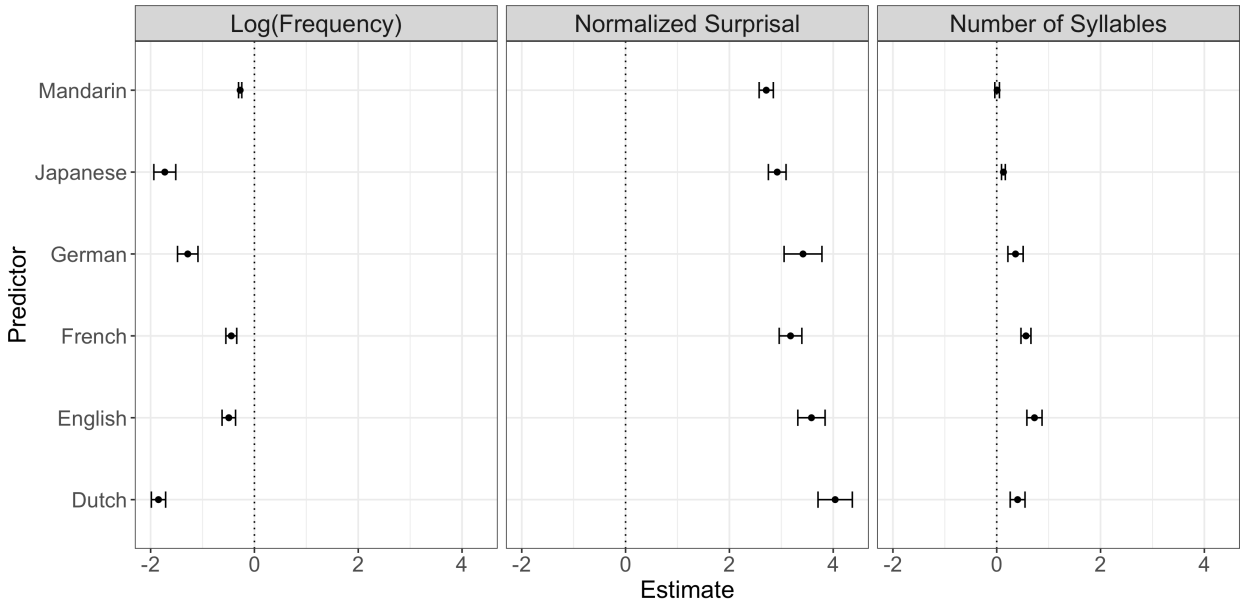


Figure 32: Parameter estimates of Homophony Delta for Log Frequency, Normalized Phonotactic Surprisal, and Number of Syllables across all six languages. Importantly, the estimates for Log Frequency are negative for each of the six languages tested. Error bars are two standard errors.

**Homophony and Neighborhood Size.** If real lexica are indeed subject to a pressure against homophony in high frequency words, that pressure should have detectable consequences elsewhere in a language. We pursued this line of reasoning by focusing on the distribution of phonological neighborhood sizes in the real lexicon. Phonological neighbors are defined as two wordforms that can be converted into one another via a single edit, i.e., a substitution, deletion, or addition (Luce & Pisoni, 1998; Vitevitch & Luce, 1999). For example, under this definition, “pot” and “pit” would be neighbors, as would “bat” and “cat”. Previous work (Dautriche et al,

2017; Trott & Bergen, 2020) has found that real languages have *larger* neighborhoods than artificial lexica matched for their phonotactics and distribution of word lengths, despite having a *smaller* number of homophones. Trott & Bergen (2020) argue that these results could arise from a pressure to avoid homophones, combined with a pressure to use high-probability phoneme sequences. Together, these pressures could create dense pockets of phonological neighborhoods in the place of a single, high-probability wordform over-saturated with meanings. If this interpretation is correct, then the wordforms most resistant to acquiring homophones should also have larger neighborhoods—i.e., controlling for other factors, Homophony Delta should be negatively correlated with Neighborhood Size.

To test this hypothesis, we added Log Neighborhood Size as a covariate to the models described above. Even accounting for the effects of Log Frequency, Normalized Phonotactic Surprisal, and Number of Syllables, the relationship between Log Neighborhood Size and Homophony Delta was significantly negative across all six languages: English [$\beta = -1.59$, $SE = 0.08$, $p < .001$], Dutch [$\beta = -1.1$, $SE = 0.23$, $p < .001$], German [$\beta = -1.85$, $SE = 0.29$, $p < .001$], French [$\beta = -3.49$, $SE = 0.14$, $p < .001$], Japanese [$\beta = -2.5$, $SE = 0.07$, $p < .001$], and Mandarin Chinese [$\beta = -2.1$, $SE = 0.05$, $p < .001$]. Wordforms with larger neighborhoods tended to have fewer homophones than predicted by their phonotactics (see also *Figure 33*). (Critically, the effect of Log Frequency remained significant across all six languages even with the addition of Log Neighborhood Size.)

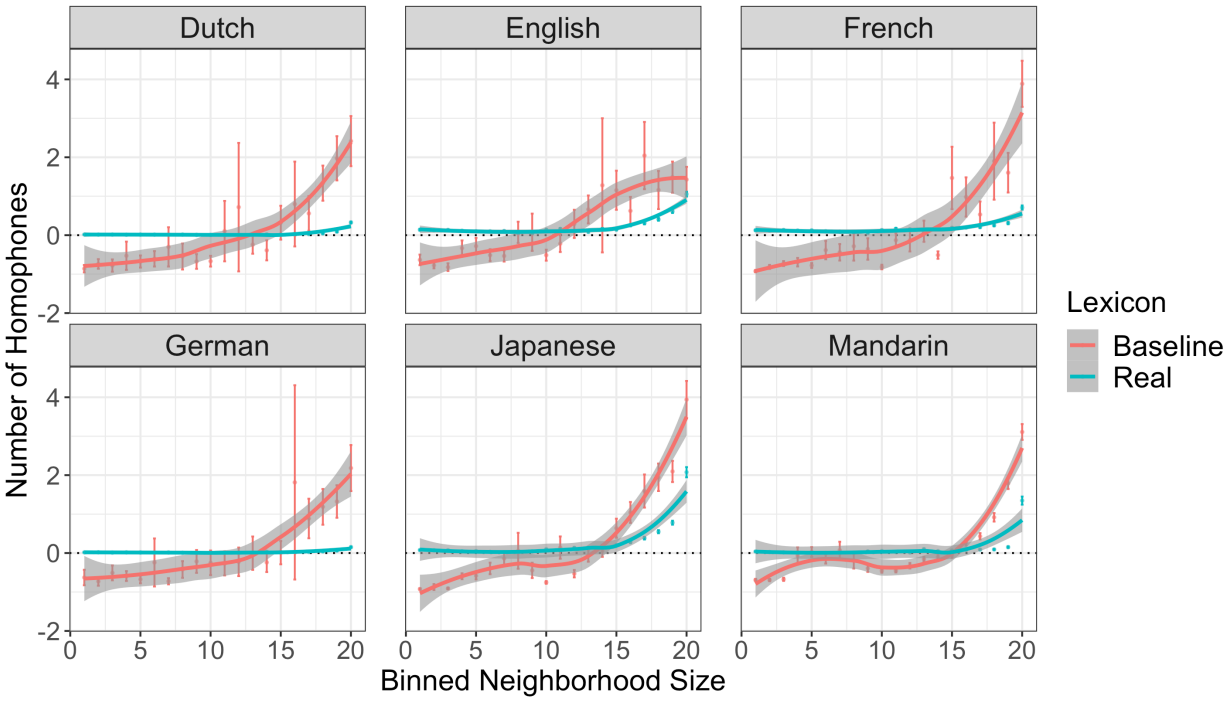


Figure 33: Real vs. predicted number of homophones, by binned neighborhood size. Wordforms with larger phonological neighborhoods tend to have more homophones in Real lexica, but this is still fewer than predicted on the basis of their phonotactics (Baseline). Higher values of Binned Neighborhood Size correspond to larger neighborhoods. Error bars are one standard error.

This relationship could be the product of a pressure to avoid homophones, which creates larger neighborhoods in their stead. But an alternate possibility exists—with reverse causality. Neighborhood size might affect the cost of disambiguation. Psycholinguistic research suggests that wordforms with larger neighborhoods are more likely to be confused with other wordforms in that language (Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Vitevitch & Luce, 1999; though see other work (Vitevitch & Rodriguez (2005); Arutiunian & Lopukhina, 2020) for evidence that this effect varies across languages). If high-density wordforms are already confusable, one might expect those wordforms to display a stronger resistance to acquiring additional meanings. On this

explanation, larger neighborhoods—like frequency—are a *cause* of a selection pressure against homophones. The current results do not allow us to adjudicate between these possibilities; however, a prediction derived from the latter interpretation is explored in the General Discussion.

## General Discussion

Our central question was the extent to which human lexica are adapted to minimize effort for speakers or comprehenders. The uneven distribution of lexical ambiguity provides a useful test case for this question: a lexicon optimized for production ease should concentrate its meanings among the easiest wordforms to produce, such as highly frequent wordforms (Zipf, 1949; Piantadosi et al, 2012). Yet such a lexicon would require frequent disambiguation on the part of comprehenders—thus, a pressure for comprehension ease would favor a lexicon with its meanings more uniformly distributed. Adjudicating between these accounts requires the use of a "neutral" baseline, i.e., a lexicon that is agnostic with respect to the relationship between wordform frequency and ambiguity and distributes its meanings according to other known factors. We used such a baseline to estimate what the magnitude of this relationship would be if meanings were assigned to wordforms with no direct pressure *for* or *against* concentrating meanings among frequent wordforms—in this case, meanings were assigned purely as a function of a wordform's phonotactic probability and length. This allowed us to compare how many meanings each wordform actually has to the number of meanings predicted by its phonotactics.

Across six languages, we found that frequent wordforms have *fewer* homophones than predicted by their phonotactics (see *Figure 31*), and in many cases, infrequent wordforms have slightly *more* homophones than expected. We also replicated this result using an alternative measure of phonotactic probability (see *Supplementary Analysis 5*). These findings are most

consistent with a Comprehender-Oriented Pressure—alternatively termed *diversification* (Zipf, 1949) or *expressivity* (Kirby et al, 2015). If each additional meaning of a wordform imposes some marginal cost for comprehenders, then a lexicon whose meanings are disproportionately concentrated among frequent wordforms will impose a larger average cost than a lexicon whose meanings are more evenly distributed. Thus, from the standpoint of minimizing comprehender effort, a selection pressure *against* homophony should manifest particularly strongly among the most frequent wordforms of a lexicon. Altogether, these results suggest that any pressure to optimize production ease is weaker than a countervailing pressure to reduce the cost of frequent disambiguation. The results of *Supplementary Analysis 6*, which formalized measures of speaker and listener effort across the lexicon, are also consistent with this conclusion. Of course, these results do not entail that human lexica are entirely shaped by comprehender-centric pressures; after all, lexica do tolerate some degree of ambiguity, even among the most frequent wordforms. Thus, a speaker-centric pressure is likely at play is well—our results simply suggest that at least when it comes to the distribution of meanings across wordforms, the comprehender-centric pressure is larger.

Further, along with other recent work (Gibson et al, 2019; Ferrer-i-Cancho et al ., 2020; Trott & Bergen, 2020; Caplan et al, 2020; Pimentel et al, 2021), these results emphasize the importance of developing formal baselines when investigating questions about the relative optimality of the lexicon.

**Limitations**

One limitation of the present work is the number and identity of languages considered. We analyzed six languages, spanning three language families (Indo-European, Japonic, and Sino-Tibetan); this sample was biased towards Indo-European languages, and did not include

languages from major families like Austronesian or Niger-Congo. We selected these languages since they are the only ones that have widely available lexical resources including information about individual meanings or lemmas, as opposed to wordforms; this was necessary for the current analyses. If similar resources become available for other languages, these analyses (and others) could be extended to a larger and more diverse set of languages.

Another potential concern is our choice of baseline, which itself might be divided into several lines of critique. The first critique is that n-gram models are prone to overfitting (see, e.g., Pimentel et al., 2021). This is a valid concern, but we have replicated the primary results using an LSTM to model phonotactics (see *Supplementary Analysis 5*), following Pimentel et al (2021). Thus, the finding that frequent wordforms have fewer homophones than predicted by their phonotactic appears robust to the phonotactic model chosen. A second concern might be that the neutral baseline is somehow not neutral—i.e., that assigning meanings to wordforms on the basis of their phonotactic probability is disproportionately biased towards speakers (or towards listeners). If this were true, it would pose a serious problem for our theoretical interpretation, which hinges on the neutrality of this assignment procedure. However, as described in the Introduction, phonotactic probability is known to facilitate both word production (Vitevitch et al, 2004; Goldrick & Larson, 2008) and word recognition and processing (Vitevitch & Luce, 1999; Vitevitch et al, 1999). To our knowledge, there is no reason to believe that phonotactic probability disproportionately benefits speakers (or listeners). This supports the neutrality of our procedure for assigning meanings to wordforms.

A third limitation or objection is that our theoretical interpretation hinges on a crucial assumption—namely, that speakers prefer a lexicon that concentrates its meanings among a few, frequent wordforms, while comprehenders prefer a lexicon that distributes its meanings more

evenly. Although this assumption is consistent with past theoretical work (Zipf, 1949; Kirby et al., 2015), we did not ground it in an explicit mathematical operationalization. Recent work (Zaslavsky et al., 2018; Zaslavsky et al., 2019; Mollica et al., 2020) has used information-theoretic tools to formalize the notions of speaker and listener effort. In *Supplementary Analysis 6*, we adopted these tools and found that, consistent with the work above, the real arrangement of wordforms and meanings is associated with lower listener effort (and higher speaker effort) than the arrangement obtained if meanings were assigned to wordforms as a function of their phonotactic probability. While this Supplementary Analysis has some limitations of its own, it is encouraging that a different methodological paradigm yielded qualitatively similar results. Future work would benefit from a more explicit grounding of the underlying semantic space.

Finally, our analyses focused on homophony. Another well-known kind of lexical ambiguity is polysemy, in which the same wordform has multiple, related meanings. Unlike homophony, polysemous words appear to enjoy advantages in both word learning (Srinivasan et al, 2019; Rodd et al, 2012; Floyd & Goldberg, 2021) and processing (Klepousniotou, 2002; Rodd et al, 2002; Klepousniotou et al, 2012). Thus, it is possible that polysemy—unlike homonymy—may even be selected for (Xu et al, 2021). If this is true, then one might also expect the opposite pattern of results to the ones reported here: the most frequent wordforms should also be *even more* polysemous than predicted by their phonotactics. In contrast, if the cost of disambiguation is still too high, a comprehender-oriented pressure for expressivity may win out even in the case of polysemy. However, one challenge to analyzing polysemy in this way is the lack of consensus about what exactly constitutes a distinct "sense" (Kilgarriff, 2007; Brown, 2008; Krishnamurthy & Nichols, 2000). Some resources make relatively fine-grained distinctions, while others aim for

more coarse-grained sense inventories (Lacerra et al, 2020). Future work in this area would thus benefit from additional resource development.

**Future Research**

Our findings point to other promising directions for future research. A first step would be to identify other factors that contribute to disambiguation cost. For example, many homophones are unbalanced, such that one meaning is used much more frequently than others. From the perspective of minimizing cost, unbalanced homophones might be preferred—if one meaning is much more frequent than another, comprehenders could simply assume the dominant meaning was intended, generally avoiding the need to disambiguate. This is consistent both with psycholinguistic evidence, which suggests that comprehenders tend to activate the dominant meaning of a homophone (Duffy et al, 1988; Blott et al, 2020), as well as work on historical sound change (Wedel et al, 2013b), which finds that phoneme mergers are especially unlikely if those mergers would create homophones among balanced minimal pairs. This interpretation also makes a testable prediction: homophones with a more uniform distribution over meanings should be more resistant to acquiring additional meanings. We tested this prediction in a supplementary analysis (see *Supplementary Analysis 2*), operationalizing meaning uncertainty as the Shannon entropy over possible senses of a wordform (Meylan et al, 2021). We found no significantly negative relationship between Sense Entropy and Homophony Delta in three of the five languages tested, though we did find a significantly negative relationship in German and Mandarin.

One explanation for these results is that disambiguation cost is driven primarily by *contextual* discriminability—i.e., how much information context provides about the intended

meaning of an ambiguous wordform. In other words, the critical factor may not be the entropy over meanings in isolation, *H(X),* but the conditional entropy over meanings given some informative context, *H(X | C)* (Piantadosi et al, 2012). Presumably, the homophones that *do* persist in a lexicon are those whose distinct meanings are sufficiently distinguishable in context (Piantadosi et al, 2012; Dautriche et al, 2018). Human comprehenders exploit a number of contextual cues to disambiguate, including grammatical class (Dautriche et al, 2018), co-speech gesture (Holle & Gunter, 2007; Holler & Beattie, 2003), linguistic context (Aina et al, 2019), and even the speaker's accent (Cai et al, 2017). Contextual discriminability should reduce the cost of disambiguation for a given wordform, thus easing the selection pressure against adding more meanings to that wordform. If this is true, a pressure against homophony should be *weaker* among wordforms with more contextually discriminable meanings, and *stronger* among wordforms whose meanings are less discriminable. Measuring contextual discriminability at scale is challenging, but future work could rely on sense-annotated corpora (Langone et al, 2004; Meylan et al, 2021), or use neural language models to derive an estimate of the residual uncertainty over meanings, given context (Pimentel et al, 2020).

The main findings reported above also inform accounts of how individual-level cognitive and communicative constraints produce emergent, lexicon-wide trends at longer timescales through language change. The presence of lexical ambiguity might elicit errors among language learners (Casenhiser, 2005) or adult comprehenders (Blott et al, 2020)—either because the cost of disambiguation was too high, or because they selected a meaning other than the one intended. Through a process of online, interactive repair (van Arkel et al, 2020), speakers might then use a different word (or series of words) to convey their intended meaning. Over many interactions, a population of speakers might drift towards using a different word in the first place, avoiding the

need for disambiguation or repair. This decision need not involve explicit or conscious ambiguity avoidance on the part of speakers, which is known to be challenging and rare (Ferreira, 2008; Wasow, 2015). Rather, it might reflect a form of implicit learning or routinization (Ferreira, 2019); the language production system might learn that when trying to convey meaning *m*, wordform $w_2$ (as opposed to ambiguous wordform $w_1$) is often used successfully. Correspondingly, the use of wordform $w_1$ to convey meaning *m* should eventually decrease, as an appropriate and less ambiguous substitute has been identified. In this way, failures of comprehension could drive future production decisions, which in turn shape lexicon structure.

Across longer timescales, one might look to processes like sound change, which are known to generate homophony (Ke, 2006; Sampson, 2013; Sampson, 2015), yet which also appear to be sensitive to a pressure to *avoid* homophones (Wedel, Kaplan, & Jackson, 2013a; Yin & White, 2018; Ceolin, 2020). For example, phoneme mergers are statistically less likely for pairs of phonemes that carry higher functional load, i.e., which distinguish more minimal pairs (Wedel, Kaplan, & Jackson, 2013a). Here, the findings above lead to another concrete prediction: phoneme mergers should be especially unlikely if they would create homophones among the most *frequent* wordforms of a language. In other words, a pressure for homophony avoidance should be strongest among frequent wordforms. To our knowledge, such a prediction has not been directly tested. A second testable prediction regarding historical sound change comes from the relationship observed above between neighborhood size and homophone resistance. One interpretation of this finding is that high-density wordforms are more perceptually confusable, and thus display a stronger resistance to acquiring more meanings. If perceptual confusability plays a role in homophone avoidance during historical sound change, phoneme mergers should be less likely if they would create homophones among high-density

wordforms. Both predictions could be tested using historical data about phoneme mergers across time and languages (Wedel et al, 2013a; Wedel et al, 2013b).

## Conclusion

Overall, our results are consistent with the claim that languages are well-designed for human use (Piantadosi et al, 2009, Gibson et al, 2019; Mahowald et al, 2020): lexica distribute their meanings in a way that reduces the cost of disambiguation. But they also support a nuanced view of "efficiency". As others (Zipf, 1949; Piantadosi et al, 2012) have noted, minimizing the effort of certain processes (e.g., production) can make other processes more challenging (e.g., disambiguation). Humans have limited cognitive resources at their disposal (Lieder & Griffiths, 2020), and these limitations create trade-offs across many domains of communication. Identifying these tension points allows us to ask more targeted questions about how this pressure for efficiency operates within and across languages. Thus, when we assess the claim that language is efficient, we might do well to begin by asking: efficient for whom?

**Acknowledgments**

Chapter 7, in full, is a reprint of the material as it will appear in the Proceedings of the 44th Annual Conference in Cognitive Science. Trott, Sean; Bergen, Benjamin (2022). The dissertation author was the primary investigator and author of this paper.

# CHAPTER 8: CONCLUSION

This thesis attempted to address two questions about lexical ambiguity. First, are word meanings categorical or continuous? The results of Chapters 2-4 support a hybrid model, in which word meanings occupy a continuous state-space (Elman, 2009), which is further discretized along the boundaries of distinct senses. And second, does the amount and distribution of homophony in real lexica reflect a pressure to concentrate meanings in the most efficient, optimal wordforms? The results suggest that homophony can emerge without a direct pressure for efficiency—and further, that real lexica might select *against* homophones, particularly among the most frequent wordforms of a lexicon. This pressure could even explain other properties of human lexica, such as their large phonological neighborhoods.

This work has limitations, and also raises additional research questions. These issues are explored at some length in Chapter 4 and Chapter 7, but I provide a brief summary in the sections below.

## Limitations

### Related topic, distinct premises

The research questions addressed in this dissertation involve a related topic (lexical ambiguity), but recruit substantively different methodologies and make distinct theoretical assumptions. Chapters 2-4 focus on how meanings are represented in the mind, and thus foregrounds the issue of meaning itself: what it looks like, where it comes from, and how it changes across different contexts. The work in chapters 5-7 focuses on the arrangement of wordforms and meanings across the lexicon, but in contrast to Chapters 2-4, does not engage deeply with the nature of meaning itself. Further, the work in chapters 5-7 assumes that word meanings are fully atomic and discrete, but the results of chapters 2-4 suggest that this

assumption is unjustified: word meaning is at least partially continuous and context-dependent. This issue is discussed at more length in the section below.

**Flawed assumptions about word meaning**

The work in chapters 5-7 makes several assumptions about word meanings, each of which is flawed or limited in some way. Some of these assumptions are unlikely to bear directly on the results, but others might; I discuss both kinds of assumptions.

First, meaning itself is modeled simply as "slots" that must be filled by phonotactically plausible wordforms; "ambiguity" is thus the number of slots filled by a particular wordform. The actual content of those slots is not considered. While this is clearly a simplification, it is worth noting that the same assumption is made by the original research to which this work is responding (Zipf, 1949; Piantadosi et al., 2012; Dautriche et al., 2017). For example, even though Piantadosi et al. (2012) consider different kinds of ambiguity (homophony and polysemy), the content of those meanings is not factored into their analyses; rather, they focus on the *number* of distinct meanings assigned to a particular wordform, as judged from a corpus like CELEX or WordNet. Second, and relatedly, these meanings are assumed to be atomic and fully distinct. The fact that some meanings relate to one another (e.g., "marinated lamb" vs. "friendly lamb") is not included in our model, nor is the fact that meanings are at least partially continuous (see chapters 2-4). Again, this assumption is also made by past work (Zipf, 1949; Piantadosi et al., 2012).

Arguably, these first two assumptions—though clearly flawed—are unlikely to have direct inferential consequences for the work described in these chapters.[40] The generative models could certainly be made more sophisticated: for example, they could sample meanings from a

[40] Though they may entail *other* assumptions that have more concerning consequences; see the paragraph below.

continuous meaning-space intended to reflect the topology of the actual lexicon, rather than assuming complete independence and atomicity of meanings. However, it is unclear how this added sophistication would change the fundamental research question being asked, or how it ought to change the results. If we accept that word meanings are *also* partially categorical (see chapters 2-4), and that some meanings are more related than others (even if homonymy and polysemy are not categorically distinct), then the simplifying assumptions described above seem suitable for asking the central question: how might we *expect* these categorical, relatively unrelated meanings to be distributed across wordforms in a lexicon absent a direct pressure for efficient reuse?

A third, more concerning assumption is the *generative process* by which new wordforms are created to convey new meanings. Our phonotactic models assume that speakers generate a new wordform "from scratch" (e.g., a kind of neologism) for each meaning; some of these newly coined wordforms happen to be identical, but the generative process is assumed to be independent. In reality, speakers of human languages have a number of meaning-generating mechanisms at their disposal. They can *extend* an existing wordform to convey a metonymically or metaphorically related meaning (e.g., Animal for Meat), or use productive *morphological rules* to create a new wordform entirely from existing lexical material (e.g., "own" can be turned into "own*er*"), or even coin a multiword expression. None of these mechanisms are considered in our model. The absence of polysemous relations and multiword expressions—although, again, a simplification—should not bear directly on the comparison between our simulated lexica and real lexica, given that the real lexica exclude polysemy and multiword expressions as well. Thus, if the research question focuses on the composition of these lexica directly—rather than the full

spectrum of how lexical meanings are created and conveyed—then these omissions should not have inferential consequences.

The more concerning omission is the lack of derivational morphology. Unlike polysemy and multiword expressions, real lexica *do* list separate entries for derivationally related wordforms (e.g., "own" and "owner"). Further, as noted in Chapter 5, derivational morphology could be a mechanism by which homophony is avoided:

> "Morphological compositionality allows speakers to convey new meanings without coining entirely new wordforms—but it also avoids the need to reuse existing wordforms for new, unrelated meanings (i.e., homophony). Thus, compositionality represents an efficient mechanism for recycling existing lexical materials that also avoids outright ambiguity." (Trott & Bergen, 2020, pg. 8).

Critically, the extent to which this is a problem depends on the extent to which derivationally related wordforms have the same number of syllables. For meanings conveyed by wordforms of different lengths (e.g., "own" and "owner"), our neglect of derivational morphology would not be responsible for generating more homophony than observed in real lexica. This is because the two meanings under consideration could not be assigned to the same wordform, given that the baselines control for how many meanings are assigned to wordforms of different lengths. However, if two meanings are conveyed by derivationally related wordforms of the *same* length, they are both candidates for being assigned to the same wordform—i.e., the baseline might independently coin the same wordform for each meaning. Rather, if applying derivational rules were an option for the baselines, that homophonous outcome might be avoided. Thus, to the extent that derivational morphology avoids homophony in real lexica, our omission of

derivational morphology could plausibly contribute to the fact that the baselines *overestimate* homophony.

This circles back to the first assumption mentioned above: meaning itself—nor the relation between meanings—is not modeled in the baselines. Although I do not believe this assumption is in itself is a problem for the research question, addressing the lack of derivational morphology might also require addressing that first assumption, along with implementing a procedure to simulate derivational morphology. Additionally, future work could consider only the set of monomorphemic wordforms in the real lexicon, and compare a baseline matched for that distribution specifically. This comparison would be more limited in its generalizability, however, so the consideration of derivational morphology seems like an important one for future work.

## Future Work

### Word Meanings: Continuous *and* Categorical?

Hybrid Theory claims that our representation of word meanings is both continuous and categorical. Yet the theory—at least as currently described—leaves a number of issues under-specified.

First, assuming that sense boundaries are in fact cognitively "real", when and how do these sense boundaries emerge during development? For example, when do the distinct senses of "lamb" start to *behave* distinctly, i.e., eliciting behavior that cannot be explained purely on the basis of a continuous account? One possibility is that certain meanings are sufficiently distinct that they are represented categorically from the beginning. Given the results of Chapter 4, this categorical distinction is unlikely to emerge—at least not "ready-made"—from differences in the distributional contexts alone; if that were true, then a model like BERT should arguably be able

to capture the observed effect of sense boundaries. It is also unlikely to emerge purely from differences in the degree to which two meanings activate similar or different sensory modalities, as a measure of sensorimotor overlap also failed to eliminate the effect of sense boundaries. Rather, this "ready-made" account might situate the distinction in properties of the referents themselves: for example, living animals and the meat produced from those animals—though related—are clearly distinct in a number of ways.[41] Our measure of sensorimotor overlap focused on the perceptual modalities themselves (e.g., vision vs. audition), but did not capture differences in the *content* of those perceptual experiences (e.g., color, shape, volume texture, etc.). It is possible that a finer-grained measure of perceptual and motor content (e.g., the 65-dimensional "brain-based vectors"—Binder et al., 2016) would provide more explanatory power; such a measure would need to be contextualized, as with the CS Norms (Chapter 3). Alternatively, one might look to representations from multimodal models, such as VL-BERT (Su et al., 2019) or DALLE-2 (Ramesh et al., 2022).

A competing possibility is that two meanings begin in one continuous cluster, but that over time, distinct sense-clusters emerge. Repeated access to distinct focal points within this continuous cluster—whether in comprehension, production, or even memory—might catalyze a process of routinization, in which the distinct meanings eventually "drift" apart. Experimental work (Srinivasan & Snedeker, 2011) suggests that in young children, related polysemous meanings share a "common representational basis", which is more consistent with this latter account. Of course, both accounts might be true, but simply for different words. In fact, these two explanations correspond roughly to the coarse taxonomy of homonymy (distinct meanings

[41] Presumably, such an account could also center the learner's own experience with these referents. This would explain why some children are surprised to learn that the lamb on their dinner plate is "the same" as the lamb at the petting zoo—and why others, perhaps with more exposure to animal husbandary, are not.

"ready-made", which share a wordform by accident) and polysemy (related meanings that "drift apart"). If both processes end in the same place (i.e., distinct senses), it would explain why the two kinds of ambiguity do not elicit reliably distinct behavior in adults.[42] Investigating these ontogenetic processes will require a mixture of methods, including: corpus of analysis of child-directed speech (Meylan et al., 2021), behavioral experiments (Srinivasan & Snedeker, 2011), and computational modeling. Modeling could adopt an approach analogous to "starting small" (Elman, 1993), and ask whether particular temporal patterns of input to a neural network can result in the emergence of sense boundaries.

Second, what exactly does it *mean* to say that meanings are continuous, or categorical, or both? Here, it is important to note that Hybrid Theory as currently described is a psychological-level theory. Further, the metaphors for word meaning that Chapters 2-4 considered—e.g., as entries in a dictionary, or as trajectories in a continuous state-space—are all under-constrained in terms of their neural implementations. On the one hand, this is not necessarily a problem for these theories: reduction to the neurobiological level of explanation is not the sole goal of psychological theory development, and neurobiological theories are not intrinsically superior in terms of their predictive or explanatory power. On the other hand, considering the brain can help *ground* debates or disagreements about psychological theory; in some cases, a commitment to particular architectural implementations of a theory can even yield useful, testable predictions at the level of behavior. From this perspective, it is easier to imagine how the mechanics of a continuous state-space could be "translated" into neural dynamics, given that the original inspiration for the state-space model was a recurrent neural network (Elman, 2009). The primary challenges, then, are: 1) enumerating the plausible architectural pathways along which these

[42] It would also mirror the distinct diachronic accounts of how homonyms and polysemes come to be.

continuous meanings are "enacted"; and 2) determining the mechanisms by which category boundaries emerge or are induced from these architectures. Here, candidate theories could be informed (and constrained) by research on other areas in which both categorical and continuous representations seem to be at play. For example, work on the categorical perception of speech (Chang et al., 2010) has found evidence for category structure in the posterior superior temporal gyrus, consistent with behavioral evidence for categorical representations of distinct phonemes. Of course, word meaning will be even more challenging to ground in the brain, given its distributed nature. Nonetheless, this work provides a useful template for thinking about categorical responses can emerge in neural behavior.

Hybrid Theory also raises questions for Natural Language Processing (NLP). As Chapters 2-4 demonstrate, current state-of-the-art NLP models are unable to fully account for human behavior on various tasks, such as relatedness judgments and primed sensibility judgments. If computational models of word meaning are meant to be humanlike—either in their mechanisms, or in their behavior—then they may benefit from the explicit representation of sense boundaries. Some such models have already been created, e.g., SenseBERT (Levine et al., 2019), using relatively coarse "super-sense" labels (e.g., *artifact* vs. *food* vs. *animal*). One question is whether training on these super-senses is sufficient to account for the apparent effect of sense boundaries; a further question is whether a training regime can be devised such that these super-senses emerge naturally, without needing explicit labels in the training data.

Finally, there is a clear connection between the work in Chapters 2-4 and debates about whether, and to what extent, cognitive processes *more generally* are best described as categorical or continuous. For example, Michael Spivey and others (Spivey, 2008; Spivey & Dale, 2006) have demonstrated that in a number of domains, apparently categorical behavior (e.g., word

recognition, semantic categorization, etc.) can emerge from underlying continuous processes. Central to their approach is the use of more sophisticated measurements that allow researchers to track continuous response dynamics (e.g., eye-tracking), as well as computational models that yield continuous predictions (e.g., neural networks). This approach represents a genuine advance in Cognitive Science, and has shown that in many cases, one need not posit explicit categorical representations to explain human cognition. Yet to our knowledge, the approach adopted in Chapter 4 is also novel. Further, our results have important implications for research that compares continuous and categorical accounts. Rather than asking solely whether a particular behavior *can* be explained by dynamics in a continuous state-space, we used statistical model comparisons to directly compare the explanatory power of this continuous account to a categorial one. This allowed us to estimate the magnitude of each parameter (i.e., contextual distance and sense boundaries) while adjusting for the other. The conclusion, which was perhaps not wholly unsurprising, was that word meanings appear to be both continuous and categorical. Having reached this conclusion, one can then ask deeper, probing questions about this hybrid model, such as how these categorical representations form and exactly how they play a role in online word processing.

In many areas of Cognitive Science, entrenched debates are sometimes cast as either/or, e.g., "Is this process continuous or categorical?" In reality, both positions in these debates are often somewhat correct—but because the question is framed in terms of binary opposition, it is unlikely to be resolved definitively either way. One benefit of the model comparison approach is that it allowed us to reframe the debate into a *parameter estimation* problem: "how much of this process is continuous, and how much is categorical?" This sidesteps the need to resolve definitively in terms of one position or the other. Further, it opens the door to more targeted,

nuanced questions, such as whether the relative magnitude of these parameters changes across task contexts, individuals, and more.

**Ambiguity: Selection Pressures and Spandrels**

The results of Chapters 5-7 suggest that contrary to previous claims (Piantadosi et al., 2012), homophones may be actively selected *against*: real lexica have fewer homophones than expected, particularly among frequent wordforms. This is interpreted as a comprehender-centric pressure to ease the cost of frequent disambiguation.

Yet as with Hybrid Theory, these results leave questions of mechanism under-specified. How would such a selection pressure be implemented? Such a question can and should be addressed at multiple levels of analysis. First, at the local level, how exactly do the challenges of disambiguation prevent wordforms from becoming over-saturated with meanings? As discussed in Chapter 7, this mechanism could manifest as repeated errors made by comprehenders in the presence of ambiguity. For example, if a speaker attaches a new meaning to an already ambiguous (but phonotactically optimal) wordform, comprehenders may struggle to resolve the intended meaning, resulting in a communication failure; if this process plays out enough times, speakers may simply choose to use a different wordform, eventually routnizing particular lexical selection processes that avoid ambiguity. Second, at the diachronic level, we might expect frequent wordforms to exhibit particular resistance to homophony-inducing sound changes; similarly, if a wordform is already ambiguous, it should be less likely to be borrowed (for a separate meaning) from other languages, relative to what one might expect given the wordform's phonotactics or the new meaning's communicative utility. Both lenses of analysis could also be applied to issues relating to the *consequences* of this selection, such as the possibility that it results in larger phonological neighborhoods (see Chapter 6).

Another set of questions concerns the alleged distinction between homonymy and polysemy. Chapters 5-7 considered only homonyms; should we expect polysemous meanings to be subject to the same selection pressures? Note that even if these phenomena are (arguably) not *psychologically* distinct, as the results of Chapters 4 suggest, they could very well be subject to distinct selection pressures, given their distinct causes. Further, the precise finding in Chapter 4 was that homonyms are not psychologically distinct from polysemes once accounting for semantic distance. Thus, even if these categories are not "real" in a psychological sense, one still might expect selection pressures for or against ambiguous words to operate differently—at least in a graded sense—according to the relatedness of their meanings. Indeed, there is some evidence that related meanings are easier to learn than unrelated meanings (Rodd et al., 2012), and that related meanings may also enjoy an advantage in processing relative to unrelated meanings (Klepousniotou, 2007). Given the apparent benefits of meaning relatedness, it is possible that polysemy—unlike homonymy—may actually be selected *for*, relative to the expectations set by a language's phonotactics and distribution of word lengths.

More broadly, the results of Chapters 5-7 also reinforce the importance of using *baselines* when asking questions about selection pressures that putatively operate over language systems. There is considerable interest in the view of language change as a kind of evolutionary process, subject to competing selection pressures that shape its trajectory through a "fitness landscape" (Kirby et al., 2015; Gibson et al., 2017). Teleological explanations are tempting in this context, particularly when one observes the presence of some widespread feature of language that appears to confer a benefit on its speakers.[43] But the conceptual challenges inherent in adaptationist accounts are well-documented for both biological (Gould & Lewontin, 1979) and cultural

[43] Voltaire satirizes this view in *Candide*: "Everything is made for the best purpose. Our noses were made to carry spectacles, so we have spectacles. Legs were clearly intended for breeches, and we wear them."

(Caplan et al., 2014) systems. The mere presence of an apparently beneficial trait does not entail that this trait was directly selected for; in complex systems—like genomes, or languages—many features emerge as the byproduct of other selection processes or even random genetic drift. As Gould & Lewontin (1979) note, the challenge is further compounded by definitional issues—determining the "conceptual boundaries" of a trait is no easy task, yet it has deep implications for which explanations we favor. With all this in mind, it is critical to consider a range of causal pathways that could give rise to a particular empirical observation (e.g., a non-zero correlation between word length and homophony); this could include the possibility of direct selection pressures (e.g., short wordforms are recycled for multiple homophones to make language more efficient), but should also include more indirect explanations (e.g., by virtue of their length, short wordforms are more likely to have more homophones simply by chance). Importantly, formal baselines are crucial for quantifying the *expectations* of these various accounts. One can then determine which account is most compatible with the empirical data, and whether direct selection pressures are *necessary* to posit.

Of course, the ability of a baseline to reproduce some empirical observation does not therefore entail that the baseline is the only correct explanation for those facts. There might still be other reasons to prefer a theory that posits direct selection: perhaps it links more closely to mechanisms of language change, or perhaps it also explains other facts about language. In this sense, one should remain clear-sighted about the inferential limits of a baseline. However, formal baselines can establish the *sufficiency* of a particular account. At the very minimum, a "successful" baseline should introduce additional uncertainty into one's beliefs about which theory is most likely true. Theory adjudication requires identifying which facts are consistent or

inconsistent with the predictions of different theories. But mere consistency with the facts is less convincing when other—perhaps simpler—theories display comparable consistency.

# REFERENCES


Aina, L., Gulordava, K., & Boleda, G. (2019). Putting words in context: LSTM language models and lexical ambiguity. arXiv preprint arXiv:1906.05149.

Alonso, D., Etienne, R. S., & McKane, A. J. (2006). The merits of neutral theory. Trends in ecology & evolution, 21(8), 451-457.

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. First Language, 40(5-6), 509-559.

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. Topics in cognitive science, 6(3), 359–370.

Armendariz, C. S., Purver, M., Pollak, S., Ljubešić, N., Ulčar, M., Vulić, I., & Pilehvar, M. T. (2020, December). SemEval-2020 task 3: Graded word similarity in context. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 36-49).

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 30, No. 30).

Arutiunian, V., & Lopukhina, A. (2020). The effects of phonological neighborhood density in childhood word production and recognition in Russian are opposite to English. Journal of Child Language, 47(6), 1244-1262.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). Distributed by the Linguistic Data Consortium, University of Pennsylvania.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? Journal of Memory and Language, 44(4), 568–591.

Bambini, V., Bertini, C., Schaeken, W., Stella, A., & Di Russo, F. (2016). Disentangling metaphor from context: an ERP study. Frontiers in psychology, 7, 559.

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.

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and brain sciences, 22(4), 577-660.

Barsalou, L. W. (2008). Grounded Cognition. Annual Review of Psychology, 59(1), 617–645.

Bates, B., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. Adaptive behavior, 11(4), 209-243.

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185-5198).

Bentley, R. A., Carrignon, S., Ruck, D. J., Valverde, S., & O'Brien, M. J. (2021). Neutral models are a tool, not a syndrome. Nature Human Behaviour, 1-2.

Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics (pp. 1-4). University of Tübingen.

Bergen, B. (2015). Embodiment, simulation and meaning. In The Routledge handbook of semantics (pp. 158-173). Routledge.

Bergen, B. K. (2012). Louder Than Words: The New Science of How the Mind Makes Meaning. New York, NY, USA: Basic Books.

Bergen, B., & Feldman, J. (2008, January). 16 - Embodied Concept Learning. In P. Calvo & A. Gomila (Eds.), Handbook of Cognitive Science (pp. 313–331). San Diego: Elsevier.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. Cognitive neuropsychology, 33(34), 130–174.

Binder, K. S., & Rayner, K. (1998). Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading. Psychonomic Bulletin & Review, 5(2), 271–276.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., Turian, J. (2020). Experience grounds language. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 8718–8735).

Blasi, D. E., Wichmann, S., Hammarstrom, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Sciences, 113(39), 10818– 10823.

Blott, L. M., Rodd, J. M., Ferreira, F., & Warren, J. E. (2020). Recovery from misinterpretations during online sentence processing. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Boleda, G., Gulordava, K., & Aina, L. (2019). Putting words in context: LSTM language models and lexical ambiguity. In Proceedings of the 57th annual meeting of the association for computational linguistics; 2019 jul 28-aug 2; Florence, Italy. Stroudsburg (pa): Acl; 2019. p. 3342–8.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. Physics of life reviews, 29, 120–153.

Boutonnet, B., Dering, B., Viñas-Guasch, N., & Thierry, G. (2013). Seeing objects through the language glass. Journal of Cognitive Neuroscience, 25(10), 1702-1710.

Brown, S. (2008). Polysemy in the Mental Lexicon. Colorado Research in Linguistics, 21(1), 2.

Brown, S. W. (2008, June). Choosing sense distinctions for WSD: Psycholinguistic evidence. In Proceedings of ACL-08: HLT, Short Papers (pp. 249-252).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. Journal of Artificial Intelligence Research, 49, 1–47.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. Current Directions in Psychological Science, 27(1), 45-50.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 dutch words. Acta psychologica, 150, 80–84.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods, 46(3), 904–911.

Burnham, K. P., & Anderson, D. R. (2002). Avoiding Pitfalls When Using Information-Theoretic Methods. The Journal of Wildlife Management, 66(3), 912–918.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. Behavioral Ecology and Sociobiology, 65(1), 23–35. https://doi.org/10.1007/s00265-010-1029-6

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. PloS one, 5(6), e10729.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. Cognitive Psychology, 98, 73-101.

Calhoon, S. W. (1935). Influence of syllabic length and rate of auditory presentation on ability to reproduce disconnected word lists. Journal of Experimental Psychology, 18(5), 612.

Caplan, S., Kodner, J., & Yang, C. (2020). Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. Cognition, 205, 104466.

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. Journal of Child Language, 32(2), 319-343.

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. Journal of Child Language, 32(2), 319–343.

Ceolin, A. (2020). On Functional Load and its Relation to the Actuation Problem. University of Pennsylvania Working Papers in Linguistics, 26(2), 6.

Chemero, A. (2011). Radical embodied cognitive science. MIT press.

Chersoni, E., Xiang, R., Lu, Q., & Huang, C.-R. (2020, December). Automatic learning of modality exclusivity norms with crosslingual word embeddings. In Proceedings of the ninth joint conference on lexical and computational semantics (pp. 32–38). Barcelona, Spain (Online): Association for Computational Linguistics.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. Journal of Child language, 30(2), 441–469.

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.

Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497–505.

Conway, B. R., Ratnasingam, S., Jara-Ettinger, J., Futrell, R., & Gibson, E. (2020). Communication efficiency of color naming across languages provides a new framework for the evolution of color terms. Cognition, 195, 104086.

Coso, B., Guasch, M., Ferr´e, P., & Hinojosa, J. A. (2019). Affective and´ concreteness norms for 3,022 croatian words. Quarterly Journal of Experimental Psychology, 72(9), 2302–2312.

Cruse, D. A. (1986). Lexical semantics. Cambridge university press.

Dautriche, I. (2015). Weaving an ambiguous lexicon (Doctoral dissertation, Sorbonne Paris Cité).

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: Homophony and the distribution of learning exemplars. Language Learning and Development, 12(3), 231-251.

Dautriche, I., Fibla, L., Fievet, A. C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. Cognitive psychology, 104, 83-105.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. Cognition, 163, 128-145.

Davis, C. P., & Yee, E. (2021). Building semantic memory from embodied and distributional language experience. Wiley Interdisciplinary Reviews: Cognitive Science, 12(5), e1555.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior research methods, 47(1), 1-12.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. Behavior research methods, 47(1), 1–12.

Deane, P. D. (1988). Polysemy and cognition. Lingua, 75(4), 325-361.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. Language and cognitive processes, 5(4), 313-349.

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, 9-37.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 4171– 4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. Journal of memory and language, 27(4), 429-446.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In Perspectives on morphological organization (pp. 212-239).

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. Topics in cognitive science, 10(1), 209-224.

Elman, J. L. (2004). An alternative view of the mental lexicon. Trends in cognitive sciences, 8(7), 301-306.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cognitive science, 33(4), 547-582.

Elman, J. L. (2011). Lexical knowledge without a lexicon?. The mental lexicon, 6(1), 1-33.

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. Computational Linguistics, 39(3), 511–554.

Faruqui, M., & Dyer, C. (2015). Non-distributional word vector representations. arXiv preprint arXiv:1506.05230.

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

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. Psychology of Learning and motivation, 49, 209-246.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. Annual review of psychology, 70, 29-51.

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2020). Optimal coding and the origins of Zipfian laws. Journal of Quantitative Linguistics, 1-30. https://doi.org/10.1080/09296174.2020.1778387

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z.,

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in Linguistic Analysis.

Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 47(1), 29.

Floyd, S., Lew-Williams, C., & Goldberg, A. E. (2019). Children, more than adults, rely on similarity to access multiple meanings of words. In CogSci (pp. 309-315).

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. Cognitive Science, 44(7), e12847.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. Brain and language, 140, 1-11.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. Journal of memory and language, 42(4), 481–496.

Futrell, R., Albright, A., Graff, P., & O'Donnell, T. J. (2017). A generative model of phonotactics. Transactions of the Association for Computational Linguistics, 5, 73-86.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences, 112(33), 10336-10341.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of experimental psychology: Human perception and performance, 6(1), 110.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., ... Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform..

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. Applied psycholinguistics, 12(3), 349-367.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. Cognitive Linguistics (includes Cognitive Linguistic Bibliography), 4(3), 223-272.

Gerz, D., Vuli´c, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. arXiv preprint arXiv:1608.00869.

Gibson, E., Futrell, R., Jara-Ettinger, J, Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S., Conway, B. (2017). Color naming across languages reflects color use. Proceedings of the National Academy of Sciences, 114(40), 10785-10790.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. Trends in cognitive sciences.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. Trends in cognitive sciences, 23(5), 389-407.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. Psychological science, 24(7), 1079-1088.

Gilhooly, K. J., & Logie, R. H. (1980). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. Behavior Research Methods & Instrumentation, 12(4), 428–450.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic bulletin & review, 9(3), 558–565.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. Cognition, 107(3), 1155-1164.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. Wiley Interdisciplinary Reviews: Cognitive Science, 1(1), 69–78. https://doi.org/10.1002/wcs.26

Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. In Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018) (pp. 10-18).

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the royal society of London. Series B. Biological Sciences, 205(1161), 581-598.

Graded word similarity in context. In Proceedings of the fourteenth workshop on semantic evaluation (pp. 36–49).

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493-498.

Gutierrez, E. D., Levy, R., & Bergen, B. (2016). Finding non-arbitrary form-meaning systematicity using stringmetric learning for kernel regression. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 2379–2388).

Haber, J., & Poesio, M. (2020). Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In Proceedings of the ninth joint conference on lexical and computational semantics (pp. 114–124).

Haber, J., & Poesio, M. (2020). Word sense distance in human similarity judgements and contextualised word embeddings. In Proceedings of the probability and meaning conference (pp. 128–145).

Haber, J., & Poesio, M. (2021, November). Patterns of polysemy and homonymy in contextualised language models. In Findings of the association for computational linguistics: Emnlp 2021 (pp. 2663–2676). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. Proceedings of the National Academy of Sciences, 117(5), 2347-2353.

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining (pp. 1406–1414).

Hanks, P. (2000). Do word meanings exist? Computers and the Humanities, 34(1/2), 205–215.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1), 335 - 346.

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. Nature, 466(7302), 29–29.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4), 665–695.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. Behavior research methods, 45(3), 718-730.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. Journal of cognitive neuroscience, 19(7), 1175-1192.

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener?. Gesture, 3(2), 127-154.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 873–882).

Johns, B. T. (2021). Distributional social semantics: Inferring word meanings from communication patterns. Cognitive Psychology, 131, 101441.

Jones, S. D., & Brandt, S. (2019). Do children really acquire dense neighbourhoods? Journal of child language, 46(6), 1260–1273.

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. Cognitive science, 44(1), e12812.

Jurafsky, D. (2014). The language of food: A linguist reads the menu. WW Norton & Company.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. Journal of memory and Language, 33(5), 630.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … Amodei, D. (2020). Scaling Laws for Neural Language Models. In arXiv.

Karidi, T., Zhou, Y., Schneider, N., Abend, O., & Srikumar, V. (2021). Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords. arXiv preprint arXiv:2109.11491.

Karidi, T., Zhou, Y., Schneider, N., Abend, O., & Srikumar, V. (2021, November). Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. In Proceedings of the 2021 conference on empirical methods in natural language processing (pp. 10300–10313). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ke, J. (2006). A cross-linguistic quantitative study of homophony. Journal of Quantitative Linguistics, 13(01), 129-159.

Kearns, K. (2006). Lexical Semantics. The handbook of English linguistics, 557.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science, 336(6084), 1049-1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. Annual Review of Linguistics, 4, 109-128.

Kempson, R. M. (1977). Semantic theory. Cambridge University Press.

Kennington, C. (2021, November). Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In Proceedings of the 25th conference on computational natural language learning (pp. 148–157). Online: Association for Computational Linguistics.

Kidd, E., & Holler, J. (2009). Children's use of gesture to resolve lexical ambiguity. Developmental Science, 12(6), 903-913.

Kilgarriff, A. (2007). Word senses. In Word Sense Disambiguation (pp. 29-46). Springer, Dordrecht.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirby, J. (2021). Incorporating tone in the calculation of phonotactic probability. In Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology (pp. 32–38).

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

Kiros, J., Chan, W., & Hinton, G. (2018, July). Illustrative language understanding: Large-scale visual grounding with image search. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 922–933). Melbourne, Australia: Association for Computational Linguistics.

Klein, D. E., & Murphy, G. L. (2001). The Representation of Polysemous Words. Journal of Memory and Language, 45(2), 259–282. https://doi.org/10.1006/jmla.2001.2779

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: conceptual relations of polysemous senses. Journal of Memory and Language, 47, 548–570.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. Brain and Language, 81(1–3), 205–223. https://doi.org/10.1006/brln.2001.2518

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. Journal of Neurolinguistics, 20(1), 1–24.

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. Brain and language, 123(1), 11-21.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6), 1534.

Krishnamurthy, R., & Nicholls, D. (2000). Peeling an Onion: The Lexicographer's Experience of Manual Sense-Tagging. Computers and the Humanities, 34(1), 85-97.

Kruyt, J. G., & Dutilh, M. W. F. (1997). A 38 million words Dutch text corpus and its users. Lexikos, 7, 229-244.

Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Perception & psychophysics, 50(2), 93-107.

Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. Working papers in corpus-based linguistics and language education, 3, 53-59.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower Perplexity is Not Always Human-Like. arXiv preprint arXiv:2106.01229.

Lacerra, C., Bevilacqua, M., Pasini, T., & Navigli, R. (2020, April). CSI: A coarse sense inventory for 85% word sense disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8123-8130).

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. Psychological Review.

Langone, H., Haskell, B. R., & Miller, G. A. (2004). Annotating wordnet. Princeton University, NJ, Cognitive Science Lab.

Lawrence W. Barsalou (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(4):577– 660. Publisher: Cambridge University Press.

Lawrence W. Barsalou (2008). Grounded Cognition. Annual Review of Psychology, 59(1):617–645.

Leinenger, M., & Rayner, K. (2013). Eye movements while reading biased homographs: Effects of prior encounter and biasing context on reducing the subordinate bias effect. Journal of Cognitive Psychology, 25(6), 665–681.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. Italian journal of linguistics, 20(1), 1-31.

Leroi, A. M., Lambert, B., Rosindell, J., Zhang, X., & Kokkoris, G. D. (2020). Neutral syndrome. Nature human behaviour, 4(8), 780-790.

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., ... Shoham, Y. (2020, July). SenseBERT: Driving some sense into BERT. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 4656–4667).

Levinson, S. (2000). Presumptive meanings: The theory of generalized conversational implicature. MIT Press

Levshina, N., & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. Linguistics Vanguard, 7(s3).

Li, J., & Joanisse, M. (2021). Word Senses as Clusters of Meaning Modulations: A Computational Model of Polysemy. Cognitive Science, 66(9), 639–646.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of experimental psychology, 54(5), 358.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behavioral and Brain Sciences, 43.

Lopukhina, A., Laurinavichyute, A., Lopukhin, K., & Dragoy, O. (2018). The mental representation of polysemy across word classes. Frontiers in psychology, 9, 192.

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2020). Language models and word sense disambiguation: An overview and analysis. arXiv preprint arXiv:2008.11608.

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

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. Behavior Research Methods, 1–21.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. Frontiers in psychology, 4, 226.

MacDonald, M. C. (2015). The emergence of language comprehension. In B. MacWhinney & W. O'Grady (Eds.), The handbook of language emergence (pp. 81–99). Chichester, UK: Wiley.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. Cognitive science, 42(8), 3116-3134.

Mahowald, K., Kachergis, G., & Frank, M. C. (2020). What counts as an exemplar model, anyway? A commentary on Ambridge (2020). First Language, 40(5-6), 608-611.

Majid, A. (2020). Human olfaction at the intersection of language, culture, and biology. Trends in Cognitive Sciences.

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4), 539.

Mark Andrews, Stefan Frank, and Gabriella Vigliocco. 2014. Reconciling embodied and distributional accounts of meaning in language. Topics in cognitive science, 6(3):359–370.

Meylan, S. C., & Griffiths, T. L. (2017). Word forms-not just their lengths-are optimized for efficient communication. arXiv preprint arXiv:1703.01694.

Meylan, S., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying Lexical Ambiguity in Speech To and From English-Learning Children.

Miceli, A., Wauthia, E., Lefebvre, L., Ris, L., & Simoes Loureiro, I. (2021). Perceptual and interoceptive strength norms for 270 french words. Frontiers in Psychology, 12, 2018.

Michaelov, J. A., & Bergen, B. K. (2020). How well does surprisal explain N400 amplitude under different experimental conditions?. arXiv preprint arXiv:2010.04844.

Miikkulainen, R., & Elman, J. (1993). Sub-symbolic natural language processing: An integrated model of scripts, lexicon, and memory. MIT press.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In Human language technology: Proceedings of a workshop held at plainsboro, new jersey, march 21-24, 1993.

Mo, L., Xu, G., Kay, P., & Tan, L. H. (2011). Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. Proceedings of the National Academy of Sciences, 108(34), 14026-14030.

Mollica, F., Bacon, G., Xu, Y., Regier, T., & Kemp, C. (2020, August). Grammatical marking and the tradeoff between code length and informativeness. In CogSci.

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

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. Journal of Speech, Language, and Hearing Research.

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. arXiv preprint arXiv:2010.13057.

Navigli, R. (2009). Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2), 1-69.

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.

Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. Nature, 551(7679), 223-226.

Ogura, M., & Wang, W. S. (2006). Ambiguity and language evolution: evolution of homophones and syllable number of words.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. Quarterly Journal of Experimental Psychology, 17(4), 273-281.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8026-8037.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp) (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In The 31st Annual Meeting of the Cognitive Science Society (CogSci09) (Vol. 2582, p. 2587). Austin, TX: Cognitive Science Society.

Piantadosi, S. T., 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. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122(3), 280-291.

Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: the wordin-context dataset for evaluating context-sensitive meaning representations. arXiv preprint arXiv:1808.09121.

Pimentel, T., Maudslay, R. H., Blasi, D., & Cotterell, R. (2020). Speakers Fill Lexical Semantic Gaps with Context. arXiv preprint arXiv:2010.02172.

Pimentel, T., Meister, C., Teufel, S., & Cotterell, R. (2021). On Homophony and Renyi Entropy. arXiv preprint arXiv:2109.13766.

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021). How (Non-) Optimal is the Lexicon?. arXiv preprint arXiv:2104.14279.

Pinker, S. (1997). Words and rules in the human brain. Nature, 387(6633), 547-548.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. The Journal of the Acoustical Society of America, 81(5), 1574-1585.

Pulvermüller, F. (1999). Words in the brain's language. Behavioral and brain sciences, 22(2), 253–279.

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. Brain and language, 127(1), 86-103.

Pustejovsky, J. (2002). The Generative Lexicon. Language, 73(3), 597.

Pustejovsky, J., & Bouillon, P. (1995). Aspectual coercion and logical polysemy. Journal of semantics, 12(2), 133-162.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL (https://www.R-project.org/).

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. Memory & cognition, 14(3), 191-201.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(5), 779.

Rayner, K., Pacht, J. M., & Duffy, S. A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. Journal of memory and language, 33(4), 527–544.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. PloS one, 11(4), e0151138.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. Proceedings of the National Academy of Sciences, 104(4), 1436-1441.

Reijnierse, W. G., Burgers, C., Bolognesi, M., & Krennmayr, T. (2019). How polysemy affects concreteness ratings: the case of metaphor. Cognitive science, 43(8), e12779.

Rice, S. A. (1992). Polysemy and lexical representation: The case of three English prepositions. In Proceedings of the fourteenth annual conference of the Cognitive Science Society (pp. 89–94).

Richie, R. (2016). Functionalism in the lexicon: Where is it, and how did it get there? The Mental Lexicon, 11(3), 429–466.

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. Perspectives on Psychological Science, 1–17. https://doi.org/10.1177/1745691619885860

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. Memory & Cognition, 40(7), 1095-1108.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. Cognitive science, 28(1), 89–104.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. Journal of Memory and Language, 46(2), 245-266.

Sampson, G. (2013). A counterexample to homophony avoidance. Diachronica, 30(4), 579-591.

Sampson, G. (2015). A Chinese phonological enigma. Journal of Chinese Linguistics, 43(2), 679-691.

Schane, S. (2002). Ambiguity and Misunderstanding in the Law. T. Jefferson L. Rev., 25, 167.

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021). DWUG: A large resource of diachronic word usage graphs in four languages. arXiv preprint arXiv:2104.08540.

Schneider, N., Srikumar, V., Hwang, J. D., & Palmer, M. (2015, June). A hierarchy with, of, and for preposition supersenses. In Proceedings of The 9th Linguistic Annotation Workshop (pp. 112-123).

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The glasgow norms: Ratings of 5,500 words on nine scales. Behavior research methods, 51(3), 1258–1270.

Sinclair, J. (1987) Collins COBUILD English Language Dictionary. London: William Collins.

Soler, A. G., & Apidianaki, M. (2021). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. Transactions of the Association for Computational Linguistics (TACL).

Speed, L. J., & Brybaert, M. (2021). Dutch sensory modality norms. Behavior Research Methods, 1–13.

Spivey, M. (2008). The continuity of mind. Oxford University Press.

Spivey, M. J., & Dale, R. (2004). On the continuity of mind: toward a dynamical account of cognition. The Psychology of Learning and Motivation, 45.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. Current Directions in Psychological Science, 15(5), 207-211.

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. Lingua, 157, 124-152.

Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. Cognitive psychology, 62(4), 245-272.

Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. Journal of Experimental Psychology: General, 148(5), 926.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. Journal of Speech, Language, and Hearing Research. 44(6): 1321-1337.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. Applied Psycholinguistics, 25(2), 201–221.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. Journal of Speech, Language, and Hearing Research.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations.

Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (cld). Behavior Research Methods, 50(6), 2606–2629.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. Cognitive psychology, 54(2), 99-132.

Tamari, R., Shani, C., Hope, T., Petruck, M. R. L., Abend, O., & Shahaf, D. (2020, July). Language (re)modelling: Towards embodied language understanding. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 6268– 6281).

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. Proceedings of the National Academy of Sciences, 106(11), 4567-4570.

Thompson, B., & Lupyan, G. (2018). Automatic estimation of lexical concreteness in 77 languages. In The 40th annual conference of the cognitive science society (pp. 1122–1127).

Trott, S., & Bergen, B. (2020). Why do human languages have homophones? Cognition, 205, 104449.

Trott, S., & Bergen, B. (2021). RAW-C: Relatedness of Ambiguous Words--in Context (A New Lexical Resource for English). Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Joint International Conference on Natural Language Processing.

Trott, S., & Bergen, B. (2021). Raw-c: Relatedness of ambiguous words in context (a new lexical resource for english). In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers) (pp. 7077–7087).

Trott, S., & Bergen, B. (2021, August). RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers) (pp. 7077– 7087). Online: Association for Computational Linguistics.

Trott, S., & Bergen, B. (2022). Languages are efficient, but for whom? Cognition, 225, 105094.

Trott, S., Torrent, T. T., Chang, N., & Schneider, N. (2020). (Re) construing Meaning in NLP. In Proceedings of the 58th annual meeting of the association for computational linguistics (acl 2020).

Trott, S., Torrent, T. T., Chang, N., & Schneider, N. (2020, July). (Re)construing meaning in NLP. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5170–5184). Online: Association for Computational Linguistics.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. Cognitive linguistics, 4(3), 273-290.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. Cognitive linguistics, 4(3), 273–290.

Turton, J., Vinson, D., & Smith, R. (2020, May). Extrapolating binder style word embeddings to new words. In Proceedings of the second workshop on linguistic and neurocognitive resources (pp. 1–8). Marseille, France: European Language Resources Association.

Uhlenbeck, E. M. (1996). Some remarks on homonymy and polysemy. Discourse and Meaning: Papers in Honor of Eva Haji? ová, 119.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. Cognitive Science, 44(6), e12844.

Valera, S. (2020). Polysemy Versus Homonymy. In Oxford Research Encyclopedia of Linguistics.

Valera, S. (2020). Polysemy versus homonymy. In Oxford research encyclopedia of linguistics.

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., & Aljasser, F. M. (2021). Phonotactics in Spoken-Word Recognition. In J.S. Pardo, L.C. Nygaard, R.E. Remez, D.B. Pisoni (Ed). The Handbook of Speech Perception (pp. 286-308). John Wiley & Sons: Hoboken, NJ, USA, 2021

Vitevitch, M. S., & Aljasser, F. M. (2021). Phonotactics in spoken-word recognition. The Handbook of Speech Perception, 286–308.

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., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. Journal of Memory and Language, 40(3), 374-408.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. Journal of Multilingual Communication Disorders, 3(1), 64-73.

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.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(2), 514.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. Brain and language, 68(1-2), 306-311.

Walker. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1):1– 48.

Wan, M., Ahrens, K., Chersoni, E., Jiang, M., Su, Q., Xiang, R., & Huang, C.-R. (2020, July). Using conceptual norms for metaphor detection. In Proceedings of the second workshop on figurative language processing (pp. 104–109). Online: Association for Computational Linguistics.

Wan, M., Xing, B., Su, Q., Liu, P., & Huang, C.-R. (2020, October). Sensorimotor enhanced neural network for metaphor detection. In Proceedings of the 34th pacific asia conference on language, information and computation (pp. 312–317). Hanoi, Vietnam: Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics.

Wasow, T. (2013). The appeal of the PDC program. Frontiers in psychology, 4, 236.

Wasow, T. (2015). Ambiguity Avoidance is Overrated. In S. Winkler (Ed.), Ambiguity: Language and communication (pp. 21–51). Berlin: DeGruyter

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. Morphology and the web of grammar: Essays in memory of Steven G. Lapointe, 265-282.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. Language and speech, 56(3), 395–417.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. Cognition, 128(2), 179-186.

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable Word Sense Disambiguation with contextualized embeddings. arXiv preprint arXiv:1909.10430.

Wingfield, C., & Connell, L. (2021). Sensorimotor distance: A fully grounded measure of semantic similarity for 800 million concept pairs.

Winter, B., & Bergen, B. (2012). Language comprehenders represent object distance both visually and auditorily. Language and Cognition, 4(1), 1–16.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. Cognition, 176, 15-30.

Wittgenstein, L. (1953.) Philosophical Investigations. Oxford: Blackwell.

Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In Proceedings of the 10th international conference on world wide web (pp. 406–414).

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). Journal of memory and language, 72, 37-48.

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 36, No. 36).

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. Cognition, 201, 104280.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5753–5763).

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. Psychonomic bulletin & review, 23(4), 1015-1027.

Yin, S. H., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. Cognition, 179, 89-101.

Yurchenko, A., Lopukhina, A., & Dragoy, O. (2020). Metaphor is between metonymy and homonymy: Evidence from event-related potentials. Frontiers in Psychology, 11, 2113.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. Proceedings of the National Academy of Sciences, 115(31), 7937–7942.

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. arXiv preprint arXiv:1905.04562.

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021). PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World. arXiv preprint arXiv:2106.00188.

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021, August). PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers) (pp. 2040– 2050).

Zipf, G. (1949). Human behavior and the principle of least effort. New York: Addison-Wesley.

Zipf, G. K. (1945). The meaning-frequency relationship of words. The Journal of general psychology, 33(2), 251-256.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley Press.

van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020, November). Explaining the efficiency of communication: How communicators can reduce their computational burden through interaction. In Proceedings of the 24th Conference on Computational Natural Language Learning (pp. 177-194).

van Esch, D. (2012). Leiden weibo corpus. http://lwc.daanvanesch.nl.