UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Syntactic distributional information in the lexicon: Systematicity, functional pressures, and grammatical implications

Permalink

https://escholarship.org/uc/item/0pn4v8zf

Author

Rogers, Phillip Gordon

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Syntactic distributional information in the lexicon: Systematicity, functional pressures, and grammatical implications

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Linguistics

by

Phillip Gordon Rogers

Committee in charge:

Professor Stefan Th. Gries, Chair
Professor Simon Todd
Professor Matthew Gordon
Professor Kyle Mahowald

June 2022

The Dissertation of Phillip Gordon Rogers is approved.

Professor Simon Todd

Professor Matthew Gordon

Professor Kyle Mahowald

Professor Stefan Th. Gries, Committee Chair

March 2022

Syntactic distributional information in the lexicon: Systematicity, functional pressures, and grammatical implications

Copyright © 2022

by

Phillip Gordon Rogers

For Luca and Zephaniah. I can't wait to meet you.

Acknowledgements

I'd first like to acknowledge the members of my dissertation committee starting with its chair, Dr. Stefan Th. Gries. You teach and advise with a candor that has been beneficial to my intellectual, professional, and personal development, and you have been an advocate for me through challenging times. I could not have asked for a better advisor. Thank you also to Drs. Simon Todd, Matthew Gordon, and Kyle Mahowald for your patience with my process, for your insightful comments and suggestions, and for investing more time and effort into this dissertation than I had any right to expect.

I am also grateful to the rest of the Linguistics faculty at the University of California, Santa Barbara—past and present—for sharing your knowledge and supporting my continued progress, even through periods when I probably didn't deserve it. Special thanks to Drs. Marianne Mithun, Bernard Comrie, Eric Campbell, and Fermín Moscoso del Prado Martín for their mentorship along the way.

Dr. William Croft was my first mentor in Linguistics and inspired my curiosity for the wonders of language diversity. Your passion and ideas continue to influence who I am as a linguist.

The research assistants who have contributed to this dissertation over the years also deserve credit for their efforts and ideas: Sherwin Lai, Wilson Wu, Alexa Theofanidis, and Serena Mao. I thrive in collaboration, and you all provided stimulation at critical moments in the development of this project.

For financial support at different stages of this project and throughout my time in the program, I thank the Graduate Division and the Department of Linguistics at the University of California, Santa Barbara.

There are many friends and colleagues who have provided intellectual insights, informed critiques, commiseration, and encouragement. I offer a special thanks to Nicholas

Lester whose research laid the foundation for this dissertation, and who invited me to collaborate with him on a related paper that sparked my interest in this topic. Tim Zingler has never let his arithmophobia get in the way of supporting my personal and academic aspirations. I've always been able to count on Danny Hieber as a friend, motivator, and sanity check. Similarly, my appreciation goes out to Nate Sims, Chris Brendel, Yi-Yang Cheng, Brendon Yoder, Michael Fiddler, Mitchell Sances, Miguel Woods, and many others. Thank you all.

This dissertation would not have been possible without the contributions of so many family members. I get my passion for learning and discovery from my dad, and my mom has always been the first to say how proud of me she is; if I had half of her work ethic, it would have been finished a long time ago. For immeasurable moral, logistical, and financial support, I thank my sisters, brother, parents-in-law, sisters- and brothers-in-law, nieces and nephews.

Finally, my deepest gratitude is reserved for my wife, Christa, and children whose patience, support, and love have been unceasing and invaluable. You remind me that the most important work I do each day happens after I close the laptop. I love you.

Curriculum Vitæ
Phillip Gordon Rogers

Education

June 2022	Ph.D. in Linguistics, University of California, Santa Barbara.
December 2015	M.A. in Linguistics, University of New Mexico.
June 2009	B.A. in History, Ohio Dominican University

Professional Experience

2015–2021	Teaching Assistant, Department of Linguistics, University of California, Santa Barbara
2016–2019	Teaching Associate (Instructor of Record), Department of Linguistics, University of California, Santa Barbara
2014	Reader, Department of Linguistics, University of California, Santa Barbara
2011–2014	Academic Advisor, College of Arts and Sciences Advisement Center, University of New Mexico
2010–2011	Campus Visit Coordinator, Admissions and Recruitment Services, University of New Mexico
2010	Student Enrollment Associate, Enrollment Management Division, University of New Mexico

Publications

Accepted	Rogers, Phillip and Stefan Th. Gries. Grammatical gender disambiguates syntactically similar nouns. [Information-Theoretic Approaches to Explaining Linguistic Structure]. *Entropy*.
2021	Kapur, Rhea and Phillip Rogers. Modeling language evolution and feature dynamics in a realistic geographic environment. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp 788–798, Barcelona, Spain, Dec. International Committee on Computational Linguistics.
2021	Rogers, Phillip. The Phonetics of Bitur. In Kate L. Lindsey and Dineke Schokkin (Eds.), *Language Documentation & Conservation Special Publication No. 24: Phonetic fieldwork in Southern New Guinea*, pp 108–119. Honolulu: University of Hawai'i Press.
2018	Lester, Nicholas A., Sandra Auderset, and Phillip Rogers. Case inflection and the functional indeterminacy of nouns: A cross-linguistic

	analysis. In Chuck Kalish, Martina A. Rau, Xiaojin Zhu, and Timothy T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pp 2029–2034.
2015	Rogers, Phillip. *Illustrating the prototype structures of parts of speech: A multidimensional scaling analysis.* MA Thesis, University of New Mexico

Awards

2014–2018	Regent's Special Fellowship, Graduate Division, University of California, Santa Barbara
2016	Runner-up, Grad Slam, University of California, Santa Barbara
2013–2014	Foreign Language and Area Studies Fellowship, Latin American and Iberian Institute, University of New Mexico

Fields of Study

Major Fields: Corpus Linguistics, Computational Linguistics, Linguistic Typology, Language Change, Language Documentation and Description

Corpus Linguistics with Professor Stefan Gries

Computational Linguistics with Professors Stefan Gries, Simon Todd, & Fermín Moscoso del Prado Martín

Linguistic Typology with Professors Bernard Comrie, Eric Campbell, & Marianne Mithun

Language Change with Professors Bernard Comrie, Eric Campbell, & Marianne Mithun

Language Documentation and Description with Professors Eric Campbell, Marianne Mithun, & Matthew Gordon

Abstract

Syntactic distributional information in the lexicon: Systematicity, functional pressures, and grammatical implications

by

Phillip Gordon Rogers

Recent psycholinguistic research has demonstrated that our knowledge of words includes fine-grained information about the syntactic contexts in which they are likely to participate. In these studies, the syntactic distribution of a word is defined as a probability distribution of its occurrences in various dependency roles. For example, nouns may be more or less likely to serve as the subject of a verb or as the head of an adjective modifier. In contrast to the constraint-based representations of mainstream generative theories, these syntactic distributions are gradient and probabilistic, situating words within a rich, multidimensional syntactic space.

At the same time, a growing body of research has identified patterns of systematicity within and among features of the lexicon that reflect cognitive and communicative pressures on learning, memory, production, and perception. For example, the same patterns of clustering and association that are observed for lexical features are also known in the psycholinguistic literature to facilitate aspects of language acquisition and use, yet these tendencies are held in check by pressures toward distinctiveness that are crucial for perception in particular.

Using corpus data from forty-eight languages, this dissertation represents the first investigation into how syntactic distributional information is patterned in the lexicon. First, I ask how syntactic representations cluster within the multidimensional syntactic space. I find that the more frequent words have denser orthographic, semantic, and—

crucially—syntactic neighborhoods. At least in phonology, having many near neighbors is known to facilitate learning, memory, and production. By analogy, it seems as though the most frequent words are also the most syntactically optimized. Next, I ask if syntactic distributions participate in non-arbitrary relationships with other features of the lexicon, such as semantics and phonology. My analysis shows that, while the meanings of words are correlated positively with both their phonological forms and syntactic distributions, phonology and syntax do not share any significant correlation below the level of word class. This surprising result suggests new ideas for how functional pressures may be negotiated at different hierarchical levels within a particular lexical feature. Finally, I ask whether the syntactic distributions of words can shed light on the organization and function of other grammatical phenomena such as grammatical gender systems. I demonstrate that, while semantically and phonologically similar words are more likely to be grouped within genders, syntactically similar words are more likely to be distributed across genders. I interpret this result as a design feature of language, with grammatical gender serving to disambiguate syntactically similar words.

Taken as a whole, the studies within this dissertation paint a novel picture of the role of syntax within the architecture of the lexicon. In some ways syntactic representations pattern similarly to semantic and phonological representations, and yet in other ways syntax seems to have a unique role relative to these other lexical features. The syntactic distributions of words reflect a kind of functional negotiation seen elsewhere in the lexicon, exhibiting both clustering and dispersion in different domains. The balance of such design features within the lexicon lend support to the idea that language structure is evolved for efficient use.

Contents

Chapter 1

Introduction

The motivation for this dissertation stems from two key ideas that have been revealed in recent linguistic research. The first key idea is the discovery of rich syntactic information in the lexicon. Many mainstream linguistic frameworks acknowledge only categorical (constraint-based) syntactic information in the lexicon (Borer, 2005; Bresnan, 2001; Chomsky, 1995; Kay, 2013; Marantz, 1997; Pollard and Sag, 1994; Ramchand, 2008). Yet usage-based approaches suggest a much richer, probabilistic integration of words and syntactic structures based on one's experience with language (Bates and MacWhinney, 1989; Bybee, 2010; Diessel, 2015). In studies on lexical recognition, production, and acquisition, Lester and colleagues have shown that when you know a word, you also know fine-grained information about the syntactic tendencies of that word (Lester, 2018; Lester et al., 2017; Lester and Moscoso del Prado Martín, 2016). These tendencies refer to the frequency with which a word participates in the various syntactic dependency relations that hold between 'heads' and 'dependents' in Dependency Grammars (Hudson, 2007; Mel'čuk, 1988; Nivre, 2005; Tesnière, 1959). For example, some nouns are more likely to occur as the subject of a verb, while others are more likely to occur as an object. Or similarly, some nouns are frequently modified by adjectives while others are rarely mod-

ified by adjectives. In this way, the syntactic distribution of a word can be defined as a probability distribution of its occurrences across dependency roles. This syntactic representation of a word is probabilistic, and it situates words within a rich, multidimensional syntactic space.

The second key idea is that there are statistical patterns within and among features of the lexicon. These patterns are instantiated in the clustering or dispersion of words in phonological or semantic space, and in correlations between these features of words. The systematicity appears to be motivated by functional pressures on learning, memory, production, and perception (Dingemanse et al., 2005). For example, the same patterns of clustering and association that are observed across features and languages are also known in the psycholinguistic literature to facilitate aspects of language acquisition and use. Even so, these tendencies are held in check by pressures toward distinctiveness that are crucial for perception in particular. The overall structure of the lexicon reflects a balance between these opposing forces, lending support to a growing body of literature attesting to the ways in which language structure is evolved for efficient use (Christiansen and Chater, 2008; Gibson et al., 2019).

At the intersection of these two ideas is the question of how syntactic distributional information is patterned in the lexicon. How do these rich syntactic representations cluster within the multidimensional syntactic space? Do syntactic distributions participate in non-arbitrary relationships with other features of the lexicon, such as semantics and phonology? Can the syntactic distributions of words shed light on the organization and function of other grammatical phenomena such as grammatical gender systems? This dissertation represents the first attempt to address these questions concerning the role of syntax in the architecture of the lexicon. Before doing so, however, it is necessary to provide more context and support for the two key ideas motivating the dissertation.

1.1 Syntax in the lexicon

In this section, I summarize the development of ideas in the literature that have led to the understanding that the lexicon contains fine-grained syntactic representations. Traditional views on syntactic information in the lexicon provide a point of departure, and from there I trace how usage-based approaches inspired research into the distributional characteristics of words—lexical, morphological, and ultimately syntactic. I anticipate that this new conceptualization of syntax will be difficult for some readers to swallow. Therefore, this background is included to illustrate how this relatively new idea has theoretical and empirical support in a broader usage-based research program.

The traditional distinction between grammar and the lexicon stems from a need to account for regularity and irregularity in language. The former allows for theoretically infinite generation from a finite set of rules, while the latter must be memorized. While most modern linguistic theories acknowledge that lexical items must be associated with *some* kind of information about how they can (or cannot) be used in syntactic structures, there remains a reluctance in dominant frameworks to allow a richer integration between words and their syntactic structures. In modern generative theories, syntactic information in the lexicon is categorical (constraint-based), limited to rules concerning the syntactic frames in which a word can participate as a head or modifier (Borer, 2005; Bresnan, 2001; Chomsky, 1995; Kay, 2013; Marantz, 1997; Pollard and Sag, 1994; Ramchand, 2008). These theories aspire to model language competence rather than performance (Chomsky, 1965), and as such see probabilistic aspects of language use as language-external and irrelevant to linguistic theory (Stabler, 2013).

In contrast, usage-based theories of language allow for a much richer representation of syntax in the lexicon. These theories posit that all aspects of language are connected in a cognitive network (Diessel, 2015; Goldberg, 2006; Langacker, 1987). The strength

of associative links between components of the network—such as words and syntactic structures—are based on one's complex experience with them and related words and structures (Bates and MacWhinney, 1989; Bybee, 2010; Diessel, 2015). Importantly, this entails associations which are probabilistic in nature. From this perspective, words are situated in a rich, multidimensional space based on their characteristics (e.g., phonological form) and distributions (e.g., across syntactic contexts).

1.1.1 Distributional effects on processing, production, and acquisition

Consistent with the predictions of these usage-based models, there is growing evidence that the distributional characteristics of words can impact language comprehension, production, and acquisition. Generally, these effects have been tied to the diversity or typicality of distributions and quantified with information-theoretic measures such as entropy and relative entropy (Shannon, 1948).

Perhaps the simplest distributional measure of a word is its lexical context—the set of words with which it co-occurs. Words that participate in diverse lexical contexts have been shown to be recognized faster in visual lexical decision tasks (McDonald and Shillcock, 2001a, 2001b). Similarly, within a particular syntactic construction, words with lexical distributions more typical for that construction are recognized faster (Baayen et al., 2011), and these effects correlate with changes in the electrophysiological signal (Hendrix et al., 2017).

Similar effects have also been demonstrated for morphological contexts. Kostić et al. (2003) found that inflectional variants of a word are processed more quickly in a visual lexical decision task if they are more frequent (relative to other inflectional variants of the same word) and if that inflected form serves fewer syntactic functions (e.g., a case

marker that expresses both objects and temporal adverbials). Put another way, when the syntactic function of an inflected form is more ambiguous, it takes longer to process. In a related vein, Moscoso del Prado Martín et al. (2004) developed the concept of *inflectional entropy*, defined as the productivity of a stem across its inflectional variants. They found that words are processed more quickly when they have a higher inflectional entropy—i.e., when occurrences of the word are more evenly distributed across its possible inflected forms. Later research demonstrated that nouns with more typical inflectional distributions (relative to all other nouns of the same gender) are also processed more quickly (Milin et al., 2009). Finally, morphological distributional effects are not limited to processing; Baayen et al. (2006) found that words with higher inflectional entropy are learned earlier in acquisition.

1.1.2 Syntactic distributional information

The findings concerning lexical and morphological distributional effects only recently inspired similar investigations into syntactic distributions. The first of these were limited to particular syntactic constructions and relations. For example, Linzen et al. (2013) found that the distribution of verbs across argument structures has a measurable effect on the electrophysiological signature of those verbs.

Lester and Moscoso del Prado Martín (2015) took a step toward a truly syntactic representation by looking at abstract—and often discontinuous—syntactic dependency relations instead of broad syntactic frames. These dependencies refer to the asymmetric relations between 'head' and 'dependent' defined within Dependency Grammar formalisms (Hudson, 2007; Mel'čuk, 1988; Nivre, 2005; Tesnière, 1959). Drawing inspiration from the research on morphological inflectional paradigms, they investigated an analogous syntactic paradigm in English: where inflectionally rich languages express nominal

roles via case inflection, English does so via prepositional phrases and word order (e.g., subjects and objects of verbs). The authors found that English nouns with higher entropy distributions across these syntactic roles had faster response times in a visual lexical decision task.

Building on this study of nominal roles in English, Lester and Moscoso del Prado Martín (2016) expanded the scope of syntactic distributions to include *all* syntactic dependencies in which a noun participates. They also defined additional syntactic entropy measures for only the relations in which a noun participates as head (it's "structure-building potential") and likewise as dependent (its ability to be integrated into other structures). They found that nouns with higher 'as-head' entropy are produced more quickly in a bare noun picture naming task, but nouns with higher 'as-dependent' entropy are produced more slowly. In lieu of these results, the authors suggest that "flexibility is not a simple component of syntactic entities, but one that interacts with different functional domains to help or hinder processing."

Lester et al. (2017) went beyond the syntactic entropy of particular words to explore the relationship *between* words and their syntactic distributions. Following the extensive literature on lexical priming (semantic, phonological, orthographic, etc.), they set out to find an effect of syntactic priming between pairs of words. They introduced an entropy-based measure of similarity between two syntactic distributions and found that this measure correlates positively with visual lexical decision priming magnitudes. Put simply, a word is recognized more quickly when its prime is syntactically more similar.

The work of Lester and colleagues on syntactic distributions is summarized and expanded in Lester (2018), which includes chapters on the processing, production, and acquisition of nouns. Processing studies in Lester (2018) include simple lexical decision and primed lexical decision. In the first of these, Lester found that both more diverse and more prototypical nouns are met with shorter response times. The diversity effect

is attributable to both the number of syntactic contexts in which the noun occurs (categorical distribution) and to the extent to which the noun is evenly distributed across these contexts (probabilistic distribution). The primed lexical decision task is essentially the study reported in Lester et al. (2017), demonstrating that syntactically similar nouns prime each other. The results of the production studies were varied. In a bare noun picture naming task there was a weak effect of typicality, with more typical nouns recognized more slowly. A second experiment asked participants to produce *the* + NOUN, and in this study higher typicality and diversity led to faster reaction times. Finally, the chapter on acquisition demonstrates that children begin producing syntactically diverse and prototypical nouns earlier in their development.

All of these studies support the idea that the lexicon contains fine-grained, probabilistic information about the syntactic distributions of words. To summarize, the evidence shows up in studies on perception, production, and acquisition, and these effects are even observed for words in isolation, not just those presented in syntactic contexts.

1.2 Functional pressures in the lexicon

Given the presence of rich syntactic information in the lexicon, this dissertation explores the ways in which that syntactic information is structured within the lexicon, both internally and with regard to other lexical features. But why should we expect syntactic information within the lexicon to be structured in the first place? The answer can be found in a growing body of research attesting to the ways in which functional pressures on learning, memory, production, and perception may shape the lexicon in different ways. The evidence for these pressures is the presence of *systematicity*, a term borrowed from Dingemanse et al. (2005) but used more broadly here. For this dissertation, systematicity refers to a statistical pattern in a single feature (clustering or dispersion beyond chance)

or a statistical relationship between two features (contrasting with arbitrariness). The presence of systematicity is generally attributed to one of two competing forces. I will introduce these competing forces briefly, and they will be exemplified in the following subsections.

The first of these forces stems from what I will refer to as the association model. These pressures are closely related to connectionist and network models of grammar in which associated items are co-activated (e.g., Diessel, 2015). The association model predicts clumpiness in single features because the activation of a particular pathway benefits from having many closely-related/associated pathways through processes of spreading activation (Collins and Loftus, 1975; Dell, 1986). The association model also predicts positive correlations between features, as the resulting compressibility of the lexicon is useful for learning and memory.

The second, competing force stems from an information-theoretic approach to language (Gibson et al., 2013; Levy, 2008a; Shannon, 1948). This model sees language as predictive and probabilistic, and language users must work together to discriminate an intended message from possible alternatives. The information-theoretic model predicts maximal differentiation across lexical structures to avoid possible confusion. Clumpiness is dispreferred for single features, as lexical items should be maximally distinct. For relationships between features, arbitrariness or even negative correlations are predicted; if two lexical items are similar with regard to one feature, they should ideally be distinct in another to avoid the possibility of confusion.

For lexical features other than syntax, there is evidence in the lexicon for systematicity both within and between features. In the following sections, we review the literature on each of these kinds of systematicity in turn. This is important for two reasons. For one, it offers a model for how syntactic information may also be patterned within the lexicon. I explicitly link each finding on systematicity to psycholinguistic research describing

the advantages of such patterns for language users, and then describe how it motivates predictions for systematic patterning in syntactic representations by analogy. Second, this research taken as a whole helps to illustrate a broader theme running through this dissertation—namely, that the lexicon is structured for efficient use.

1.2.1 Systematicity within a single feature

Most of the research on the systematicity of a single feature of the lexicon has focused on wordforms (phonology). There are several sources of phonological regularity in the lexicon. One such source is morphology, as words that share a morpheme will also typically share its phonological form. However, there are other sources of regularity that also apply to mono-morphemic words. One of these is phonotactics—the complex set of rules concerning which sounds and sound sequences are allowed in a language (Hayes and Wilson, 2008; Vitevitch and Luce, 1998)—and the regularity introduced by phonotactics goes beyond categorical constraints. As it turns out, phonotactically probable words are recognized more quickly than less probable words (Vitevitch et al., 1999) and learned more quickly by infants and young children (Coady and Aslin, 2004; Storkel and Hoover, 2010). Similarly, infants prefer to listen to high-probability sequences over low-probability ones (Jusczyk et al., 1994; Ngon et al., 2013).

What's more, the clustering of phonological forms goes above and beyond the effects of phonotactics and morphology (Dautriche et al., 2017). Research on phonological neighborhood density effects sheds light on the advantages that these clusters offer. Words in high-density neighborhoods are produced faster and more accurately (Dell and Gordon, 2011; Gahl et al., 2012; Stemberger, 2004; Vitevitch, 2002; Vitevitch and Sommers, 2003, but see Sadat et al., 2014), and they are learned more easily by both children (Storkel, 2004) and adults (Storkel et al., 2006). In fact, similarity with known words may help in-

fants identify novel words in fluent speech (Altvater-Mackensen and Mani, 2013). There is also evidence that this kind of phonological similarity facilitates short- and long-term memory (Storkel and Lee, 2011; Vitevitch et al., 2012).

Yet another source of systematicity in the phonology of a language is homonymy, whereby multiple meanings share the same form. There is evidence that this lexical ambiguity can actually facilitate processing by permitting efficient linguistic units to be re-used (Piantadosi et al., 2012). Additionally, homonyms are produced more accurately than novel word forms by children in a word learning task (Storkel and Maekawa, 2005). Finally, experimental work on language evolution has shown how this systematic under-specification in the mapping of form to meaning can arise through language transmission (Kirby et al., 2008).

Systematicity between features (e.g., between phonology and semantics) are the source of additional phonological regularity, but these relationships are addressed in the next section. To summarize the literature reviewed so far, it appears that lexicons with high phonological systematicity offer advantages to learning, memory, and production (Monaghan et al., 2011).

However, there is also evidence that phonological regularity can be detrimental, particularly for perception. Words with more phonological neighbors are processed more slowly and with more errors (Luce and Pisoni, 1998; Vitevitch and Luce, 1998; Vitevitch et al., 1999). These inhibitory effects on comprehension can also negatively effect a toddler's ability to learn novel words with close phonological neighbors (Swingley and Aslin, 2007). These findings represent a pressure for dispersion which is predicted by the information-theoretic model introduced above. Perceptual distinctiveness is crucial for guarding the lexicon against noise, and it is a key design feature of phonological systems (Flemming, 2004; Graff, 2012; Lindblom, 1986; Wedel et al., 2013). Other studies have shown how language production is modulated to preserve perceptual distinctions, reflect-

ing a sensitivity to factors such as frequency, predictability, and potential confusability (Aylett and Turk, 2004, 2006; Bell et al., 2003; Cohen Priva, 2008; Levy and Jaeger, 2007; Pluymaekers et al., 2005; Raymond et al., 2006; Van Son and Van Santen, 2005).

The research described in this section suggests that the degree of phonological systematicity in the lexicon reflects a balance between pressures of association and dispersion. Some regularity (clustering) aids learning, memory, and production, while distinctiveness serves perceptual needs. While this research has focused on phonology, I hypothesize that the same principles should apply to any lexical feature, including syntax. According to usage-based theories, distributional characteristics of a word such as syntax share important commonalities with other lexical features such as wordforms. For both, associations within the cognitive network are strengthened over time through one's experiences with language, and the advantages of clustering can be explained by spreading activation (Collins and Loftus, 1975; Dell, 1986). In this way, language users may stand to gain as much from systematicity in syntactic distributional information as they do from systematicity in phonological forms.

In this dissertation, I ask whether syntactic distributional information in the lexicon is subject to similar patterns of internal systematicity. Just as for wordforms, there are several ways that syntactic information could exhibit such systematicity. First, syntactic distributions might be clustered beyond what would be expected by chance. This could be measured in several ways; for example, clustering of phonological forms in real lexicons was been measured against simulated lexicons using average distance between wordforms, number of minimal pairs, and network measures (Dautriche et al., 2017). Second, one could explore the relationship between frequency and syntactic distributions. Mahowald et al. (2018) found that more frequent words are more orthographically well-formed and have denser phonological neighborhoods, interpreting this as evidence that the most frequent words are also most optimized for efficient use. One could ask similar questions

about syntactic distributions.

Short of answering all of these questions, this dissertation takes a first step to addressing them in Chapter 3. In that study, I assess the relationship between frequency and syntactic neighborhood density in nearly fifty languages. To my knowledge, it constitutes the first study of syntactic neighborhood density and the first to investigate systematicity within syntactic distributional information. My findings in that study invite additional research into other patterns of internal systematicity within syntactic distributional information.

1.2.2 Systematicity in the relationship between features

Turning now to systematicity between features of the lexicon, much of the focus has been on the relationship between form and meaning. Traditionally, this relation is considered to be arbitrary (de Saussure, 1916; Hockett, 1960), yet there are well-known examples of systematicity and growing evidence for the usefulness of such patterns (Dingemanse et al., 2005). One source of this systematicity is sound symbolism, the iconic mapping of sound to meaning. There are many different types of sound symbolism, and patterns of sound symbolism are present in many languages and cultures (Blasi et al., 2016; Bremner et al., 2013; Hinton et al., 1994; Lockwood and Dingemanse, 2015, inter alia). Yet the systematicity goes beyond what can be accounted for by sound symbolism. Several studies have demonstrated a widespread correlation between form and meaning across the lexicon (Dautriche et al., 2016; Monaghan et al., 2014; Shillcock et al., 2001; Tamariz, 2008).

These regular correspondences appear to have advantages for learning and memory. A correlation between form and meaning has been shown to facilitate learning in adults and children (Imai and Kita, 2014; Imai et al., 2008; Monaghan et al., 2014; Nielsen and

Rendall, 2012; Nygaard et al., 2009). The learning bias is further supported by Monaghan et al. (2014), who report more systematicity in words acquired earlier. Furthermore, a positive correlation between phonology and semantics results in redundancy, as some aspects of a word can be at least partially predicted by other aspects. In this way lexical knowledge is compressible (Kirby et al., 2015; Tamariz and Kirby, 2015), easing demands on learning and memory.

Phonology has also been shown to mark both semantic and syntactic categories. For example, Reilly et al. (2012) found that there are phonological correlates to the semantic distinction between concrete and abstract words. Concerning syntactic categories, several studies have drawn attention to statistical differences across a variety of phonological properties that cue grammatical distinctions such as noun vs. verb and open vs. closed word class (Cassidy and Kelly, 1991; Kelly, 1992; Monaghan and Christiansen, 2008; Monaghan et al., 2005; Monaghan et al., 2007). These phonological cues have been shown to support category learning and generalization to novel words (Fitneva et al., 2009; Monaghan et al., 2011).

However, the advantages of systematicity between features seem to be limited. While form-meaning correspondences give advantages for category learning, some studies suggest they do not facilitate acquisition of individual word meanings (Monaghan et al., 2011; Monaghan et al., 2012). For example, toddlers find it difficult to learn words that are similar in both form and meaning (Dautriche et al., 2015; Swingley and Aslin, 2007). Form-meaning regularity can also be detrimental to production. The mixed error effect describes the phenomenon by which speech error substitutions are more common for words that are both semantically and phonologically similar than for words that share only one of these properties (Dell and Reich, 1981; Goldrick and Rapp, 2002; Schwartz et al., 2006). Taken to an extreme, systematicity would lead to the presence of highly confusable words throughout the lexicon.

These findings are a reminder that arbitrariness still plays an important role in the lexicon. Several studies demonstrate that arbitrary forms are more effective than iconic forms and nonverbal cues when it comes to activating general and abstract representations (Boutonnet and Lupyan, 2015; Edmiston and Lupyan, 2015; Lupyan and Thompson-Schill, 2012). There is also evidence that iconic form-meaning mappings work well in small vocabularies, but as the size of the vocabulary increases, the potential for confusion increases and arbitrary mappings become more efficient (Gasser, 2005). Finally, arbitrary form-meaning pairs may be a prerequisite for the transition to a combinatorial language system, in which a finite number of signs and parts of signs are combined into a theoretically infinite number of utterances (Hockett, 1960; Nowak et al., 1999).

To summarize the literature reviewed here, systematicity exists to a degree within the lexicon, but it is kept at bay by opposing pressures favoring arbitrariness, with both offering advantages to different aspects of language acquisition and use. For example, Monaghan et al. (2014) suggest that the lexicon is structured such that regularity promotes early language learning while arbitrariness is introduced later to facilitate communicative expressivity and efficiency.

Again, I ask whether the syntactic distributions of words are subject to similar patterns of systematicity with regard to their relations to other features of the lexicon. In particular, I ask whether syntactic representations are correlated with wordforms and word meanings. I hypothesize that these relationships will be positive correlations, just like the relationship that has been demonstrated between form and meaning. These kinds of systematic relationships would presumably imbue the same advantages for learning and memory that are observed for form-meaning correspondences. This question is addressed in Chapter 4 of the dissertation.

1.3 Scope of the dissertation

This dissertation sets out to explore three patterns of systematicity concerning syntactic distributional information cross-linguistically. Before investigating these patterns, Chapter 2 orients the reader with the data and methods shared among the following chapters. In particular, I demonstrate how syntactic information is extracted from annotated corpora and converted into probability vectors. To make these abstract ideas more accessible to readers, some concrete examples are given of syntactic distributions and syntactic comparisons between words.

As suggested above, Chapters 3 and 4 are aimed at revealing whether syntactic distributions participate in patterns of systematicity in the lexicon. Chapter 3 focuses on regularity within syntactic distributions testing whether syntactic neighborhood density is correlated with frequency across languages. I hypothesize that high frequency words will have denser syntactic neighborhoods, on analogy with the findings concerning phonology in Mahowald et al. (2018). Chapter 4 examines systematicity between syntactic distributional information and other features of the lexicon: wordforms and word meanings. I expect positive correlations, consistent with findings on the relationship between form and meaning (Dautriche et al., 2016).

In Chapter 5, I am interested in a different kind of grammatical system that has close ties to the lexicon: grammatical gender. In that study, I investigate the relationship between syntactic distributional information and grammatical gender assignment. The literature suggests that grammatical gender assignment may be functionally motivated to disambiguate potentially confusable nouns, reminiscent of the pressures for dispersion discussed above. I propose syntax as a locus for this disambiguation, as syntactically similar words will compete for activation in those syntactic contexts that they share. To test this hypothesis, I ask whether syntactically similar words are more or less likely to

share a gender across languages.

I bring together the results of all these studies in Chapter 6 to discuss their implications for the lexicon and grammatical theory. Although these are the first studies of syntactic systematicity in the lexicon, I can draw on the results of each chapter to formulate preliminary answers to the questions I have posed. Do syntactic distributions exhibit systematicity, either internally or in their relations to other features of the lexicon? If so, do these patterns of systematicity reflect pressures of association or dispersion? Can syntactic distributional information provide insights into the function of grammatical gender systems? I conclude in this chapter with shortcomings and future directions.

Chapter 2

Data and Methods

The primary source of data for this study is the Universal Dependencies Treebanks (UDT) (de Marneffe et al., 2021). This project offers cross-linguistically consistent part of speech tagging and dependency annotation for data from over 100 languages. Corpus size and the availability of additional features such as lemmas vary from language to language, and many languages are represented by multiple corpora. I extracted wordform, lemma, part of speech, gender, and syntactic information for every token of every corpus in UDT, excluding corpora without consistent lemma information.

2.1 Syntactic representations

The syntactic information of a word consists of every syntactic dependency that the word participates in, either as a head or dependent. The UDT dependency framework is illustrated in Figure 2.1. In this example, the Spanish word *oro* ('gold') participates in two syntactic dependencies. First, it is the head of a case relation with *de* ('of/from'). Second, it is the dependent of a nominal modification relation with *medallas* ('medals'). These two relations highlight an important characteristic of the UDT framework: the

primacy of content words. Practically speaking, this means UDT dependencies link content words directly rather than indirectly through function words. In contrast, many dependency grammars would view *oro* as a dependent of *de*, which in turn would be viewed as a dependent of *medallas*. For my purposes, I am interested in the overall syntactic distributions of words, so the particular framework by which those dependencies are annotated matters less than the consistency by which that framework is applied across sentences and languages.

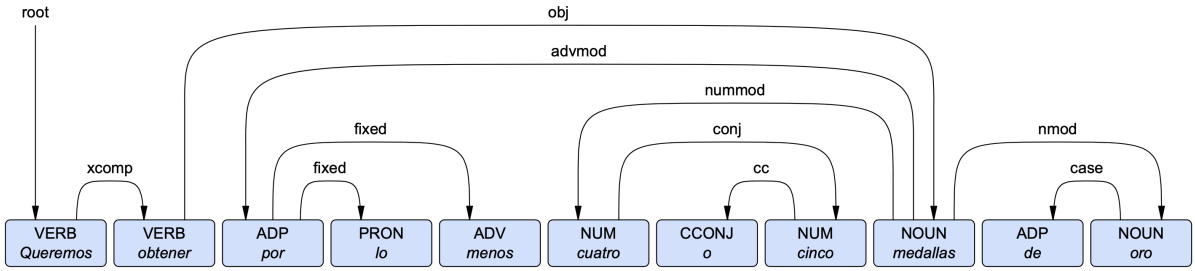

Figure 2.1: An example of the Universal Dependencies Treebanks dependency framework from Spanish. Syntactic dependencies are represented by arrows pointing from heads to their dependents, and each dependency is labeled for the type of relation. The translation of the sentence is 'We want to get at least four or five gold medals.'

The studies in this dissertation aim to address the lexicon, and therefore I aggregate the UDT data by lemma and part of speech. (Alternatively, one could aggregate tokens by wordforms within or across parts of speech.) The decision to investigate lemmas is particularly important for Chapter 5 because grammatical gender is a feature of lexemes rather than specific wordforms. Upon aggregation, syntactic information takes the form of a syntactic vector. Each position in the vector represents a specific syntactic role and relation, such as head of a determiner relation. The value at that position represents how many times a particular lemma was attested in that relation and role. As such, the entire vector constitutes a frequency distribution of the syntactic dependency types in which a lemma has participated.

Frequency distributions are known to be biased by sample size. Following Lester

(2018), I correct these distributions using the James-Stein shrinkage estimator (Hausser and Strimmer, 2009). This bias correction method performs well on data for which the number of types is known, and—given the size of the corpora—I assume that the dependency types represented by the corpus data are exhaustive. The bias correction also transforms the syntactic vector from a frequency distribution into a probability distribution.

I illustrate the syntactic distributions of lemmas with three examples from Spanish. Figure 2.2 shows partial probability vectors for three Spanish lemmas that are well attested in the data: *medalla* (n = 72), *oro* (98), and *paz* (132). The ten dependency types included in the illustration are only a subset of those found in Spanish, but they include types in which nouns often participate (they account for 87% of *medalla* dependencies, 93% of *oro*, and 96% of *paz*). The height of each bar represents the (bias-corrected) rate at which that lemma participates in that dependency type relative to other dependency types.

It is readily apparent from Figure 2 that *oro* and *paz* are much more similar to each other syntactically than either is to *medalla*. Both *oro* and *paz* participate frequently as the dependent in a nominal modifier dependency and as the head of a case marking dependency, while *medalla* does not. On the other hand, *medalla* is far more likely to occur as an object of a verb and as the head of a nominal modifier dependency. All four of these dependency types are illustrated with *oro* and *medalla(s)* in Figure 2.1. In fact, *oro* and *medalla* co-occur frequently in the corpus in the phrase *medalla(s) de oro* ('gold medal(s)'), contributing to the patterns observed in their syntactic distributions. These words—*oro* and *medalla*—are semantically similar yet syntactically distinct. In contrast, *oro* and *paz* are semantically unrelated yet syntactically similar.

I also derive measures of frequency from the UDT corpora. Token frequencies from the UDT are normalized across languages using the total corpus size for each language

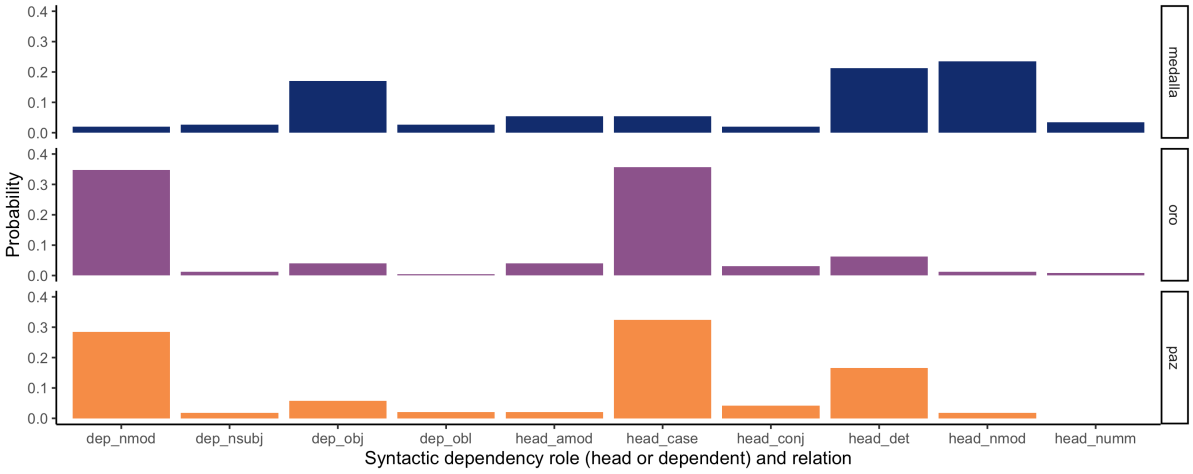

Figure 2.2: Partial probability vectors for the participation of three Spanish lemmas in different syntactic roles and relations. The height of each bar indicates how often that lemma participates in that dependency type relative to other syntactic dependency types. The probabilities shown are corrected for sample bias with the James-Stein shrinkage estimator. These three distributions illustrate how *oro* is much more similar syntactically to *paz* than to *medalla*, despite being more similar semantically to the latter.

and converted to per-million frequencies. I exclude lemmas occurring less than ten times in the data. This cut-off represents a compromise between excluding words whose syntactic distributions are unreliable and including as many words as possible. Because of the Zipfian distribution of word frequencies, increasing the minimum frequency even a little excludes many words from the analysis. On the other hand, including very low frequency lemmas may introduce unwanted biases into the analyses. For example, very low frequency lemmas are more likely to share an exact syntactic distribution with other very low frequency lemmas in the sample, obscuring subtle but important distinctions in their real syntactic distributions.

2.2 Semantic and orthographic data

While the focus of the dissertation is on syntactic distributional information, each study also incorporates semantics and orthography (as a proxy for phonology). In Chapter 4 these other word features are central to the hypotheses, while in Chapters 3 and 5 they are included as controls and to reproduce results from the literature. The UDT corpora are too small to produce reliable semantic vectors, so fastText semantic vectors (Bojanowski et al., 2016) are matched to words in the UDT. As the vectors from fastText correspond to wordforms, I compute weighted averages (by frequency) for lemmas to make them compatible with the lemma-aggregated UDT data. Similarly, phonological transcriptions are not available for the languages in the study, so I utilize orthography as a proxy for phonology. This is justifiable based on previous work: Dautriche et al. (2016) examined the relationship between phonology and orthography and found high correlation between the number of phonemes and characters in a word in Dutch (r = .87), English (r = .83), German (r = .89), and French (r = .79).

2.3 Languages

The resulting data is further trimmed and transformed for each study in the dissertation, as outlined in their respective chapters. For example, Chapter 5 excludes languages without grammatical gender. In all, 48 languages are included in at least one of the analyses, as shown in Table 2.1. I include family and subfamily affiliations for each language, and this information is used in analyses to identify effects that are specific to a language group. Groupings are based loosely on Glottolog (Hammarström et al., 2021); yet with many potential grouping levels, I had to make subjective decisions with regard to which levels to include, particularly within Indo-European. It would have been reasonable to

group all ten Slavic languages together in one large subfamily, but I chose to use East, South, and West Slavic groups instead to highlight trends within each. With regard to controversial language families, I take a conservative approach. For example, Turkish, Korean, and Japanese are assigned to their own families despite putative membership in a Transeurasian family (e.g., Robbeets, 2017). Similarly, the names chosen for families and subfamilies used in these groupings are not intended to be controversial, and they should not be interpreted beyond the distinctions they make in the data.

The approach of this dissertation is intentionally cross-linguistic, to the extent allowed by data availability. A more detailed and extensive investigation of one or a few languages was sacrificed in favor of one that included as many languages as possible. Ten families and twelve subfamilies within the Indo-European family are represented in this sample. Yet, ideally, the sample for these investigations would be far more diverse in terms of both geographic and genealogical spread (Bickel, 2008; Dryer, 1989; Miestamo et al., 2016; Rijkhoff and Bakker, 1998, inter alia). There are no indigenous languages of the Americas, Sub-Saharan Africa, or Australia. This is addressed again in Chapter 6, and the results presented throughout this dissertation will be subject to future reconsideration as corpora and other resources become available for many more languages and language families.

2.4 Distance measures

One way or another, each of the studies in this dissertation require comparisons between words with regard to their syntactic, semantic, and orthographic properties. I perform these comparisons with distance metrics which are appropriate for each lexical feature. The process of comparing words to each other has the added benefit of reducing lengthy semantic and syntactic vectors and complex orthographic strings to a distance

Language	Subfamily	Family	Ch. 3	Ch. 4	Ch. 5
Arabic	Semitic	Afro-Asiatic	1,989	252,676	NA
Hebrew	Semitic	Afro-Asiatic	1,169	110,366	332,520
Vietnamese	Mon-Khmer	Austro-Asiatic	329	9,107	NA
Indonesian	Malayo-Polynesian	Austronesian	1,172	69,176	NA
Basque	Basque	Basque	899	35,308	NA
Armenian	Armenic	Indo-European	431	7,093	NA
Latvian	Baltic	Indo-European	2,035	158,940	853,470
Lithuanian	Baltic	Indo-European	705	20,631	116,886
Gaelic	Celtic	Indo-European	333	6,597	27,261
Irish	Celtic	Indo-European	846	36,527	193,131
Welsh	Celtic	Indo-European	272	5,420	16,290
Belarusian	East Slavic	Indo-European	1,860	150,525	808,356
Russian	East Slavic	Indo-European	6,299	1,459,997	9,033,375
Ukrainian	East Slavic	Indo-European	874	36,285	203,203
Greek	Hellenic	Indo-European	485	9,466	55,611
Hindi	Indo-Aryan	Indo-European	1,547	203,945	914,566
Urdu	Indo-Aryan	Indo-European	796	73,315	263,796
Persian	Iranian	Indo-European	3,436	1,135,740	NA
Danish	North Germanic	Indo-European	539	14,040	53,300
Icelandic	North Germanic	Indo-European	3,060	435,333	1,547,919
Norwegian	North Germanic	Indo-European	2,709	305,085	1,828,828
Swedish	North Germanic	Indo-European	1,159	54,259	287,660
Catalan	Romance	Indo-European	2,345	255,197	1,189,652
French	Romance	Indo-European	2,697	305,373	1,941,435
Galician	Romance	Indo-European	1,151	48,639	NA
Italian	Romance	Indo-European	3,287	454,336	2,657,665
Latin	Romance	Indo-European	2,308	220,168	945,996
Portuguese	Romance	Indo-European	1,408	81,728	478,731
Romanian	Romance	Indo-European	3,432	472,697	3,012,284
Spanish	Romance	Indo-European	3,729	691,964	3,121,251
Bulgarian	South Slavic	Indo-European	1,139	49,548	307,720
Croatian	South Slavic	Indo-European	1,424	93,180	533,028
Serbian	South Slavic	Indo-European	762	26,114	152,628
Slovenian	South Slavic	Indo-European	1,052	53,956	268,278
Afrikaans	West Germanic	Indo-European	358	4,074	NA
Dutch	West Germanic	Indo-European	1,463	90,288	337,431
English	West Germanic	Indo-European	2,570	303,346	NA
German	West Germanic	Indo-European	6,672	2,095,196	9,660,210
Czech	West Slavic	Indo-European	7,590	3,191,510	15,326,415
Polish	West Slavic	Indo-European	3,222	429,515	2,336,041
Slovak	West Slavic	Indo-European	773	28,110	120,295
Japanese	Japonic	Japonic	1,727	13,349	NA
Korean	Koreanic	Koreanic	3,010	719,907	NA
Chinese	Sinitic	Sino-Tibetan	1,891	3,757	NA
Turkish	Turkic	Turkic	3,171	719,273	NA
Estonian	Finnic	Uralic	2,671	335,563	NA
Finnish	Finnic	Uralic	2,219	212,619	NA
Hungarian	Hungarian	Uralic	254	3,116	NA

Table 2.1: Languages, their genealogical affiliations, and the number of data points used in Chapters 3–5. Chapter 3 numbers refer to lemmas included in the analyses, while Chapters 4 and 5 refer to the number of lemma pairs.

between two vectors/strings.

For orthography, I use Levenshtein distance, defined as the minimum number of single-character insertions, deletions, and/or substitutions needed to change one character string into another.

For semantics, I use Cosine similarity, as it is the standard metric for measuring similarity between two semantic vectors. This metric is popular because it captures the angle between the vectors in multidimensional space, ignoring the magnitude of those vectors. Subtracting the cosine similarity from 1 turns it into a distance metric. Given vectors A and B, where Ai and Bi are the components of these vectors, the formula for cosine similarity is

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}. \tag{2.1}$$

Finally, for syntax, I use the entropy-based Jensen-Shannon Divergence (JSD) between syntactic vectors (following Lester et al., 2017). JSD is a bounded, symmetric distance metric based on the Kullback-Leibler Divergence (KLD). KLD is an unbounded, directional (asymmetric) measure of the information loss of approximating one probability distribution by another, and JSD makes this measure bidirectional by averaging the distance to the midpoint of the two distributions. The relevant equations for JSD are as follows for probability distributions P and Q defined on the probability space X:

$$JSD(P\|Q) = \frac{1}{2}KLD(P\|M) + \frac{1}{2}KLD(Q\|M); \tag{2.2}$$

$$KLD(P\|Q) = \sum_{x\in X} P(x) log\left(\frac{P(x)}{Q(x)}\right); \tag{2.3}$$

$$M = \frac{1}{2}(P + Q). \tag{2.4}$$

Chapter 3

Systematicity within syntactic distributions

3.1 Background

In Chapter 1, I introduced the notion of systematicity in the organization of lexicon. Systematicity has been shown to exist within phonological forms, as wordforms are clustered beyond the influence of phonotactics and morphology (Dautriche et al., 2017). This clustering offers advantages to learning (Storkel, 2004; Storkel et al., 2006), memory (Storkel and Lee, 2011; Vitevitch et al., 2012), and production (Dell and Gordon, 2011; Gahl et al., 2012; Stemberger, 2004; Vitevitch, 2002; Vitevitch and Sommers, 2003). Of particular interest for this study is the observation that the clustering of phonological forms further correlates with frequency, such that the most frequent forms are also those with the densest neighborhoods (Mahowald et al., 2018). This means that the advantages conferred by this regularity are greatest for the words that are used the most. This finding is consistent with information-theoretic accounts that predict that the most frequent words should be most optimized for efficiency (Piantadosi et al., 2011; Zipf, 1949).

(It should be noted that in Chapter 1 I discussed information theory primarily in the context of pressures of dispersion, as messages must be discriminated from possible alternatives. This was contrasted with association pressures predicting clustering. However, information theory also makes this meta-prediction about the way the lexicon will be organized with regard to frequency. My predictions about syntactic neighborhood density and frequency refer to advantages conferred by clustering (as opposed to advantages of dispersion), and yet such a finding can be reconciled with an information-theoretic framework in this way.)

The semantics landscape is less clear. For one, it is complicated by the diversity of approaches to capturing semantic information (see below). Additionally, most studies on the effects of semantic neighborhood density focus on lexical recognition and categorization tasks at the expense of learning, memory, and production (e.g., Buchanan et al., 2001; Locker et al., 2003; Mirman and Magnuson, 2008; Nelson et al., 1992; Siakaluk et al., 2003; Yates et al., 2003). What is clear from these studies is that semantic neighborhood density correlates strongly with frequency (e.g., Mirman and Magnuson, 2008), just as is observed for phonology.

I hypothesize that this same pattern of systematicity will be found with syntactic representations. I expect to find that more frequent words will have denser syntactic neighborhoods, as these associations presumably offer the same advantages to language acquisition and use as they do in phonology. To my knowledge, this study constitutes the first investigation into syntactic neighborhood phenomena in the literature.

3.2 Analysis

Data for this study were collected according to the methods described in Chapter 2. After removing lemmas that did not meet the token frequency threshold of ten,

syntactic, semantic, and orthographic neighborhood densities were computed. However, for the statistical analyses described below, the data was additionally subset to just nouns and verbs for important reasons. Syntactic distributions, by definition, reflect the part of speech of a target word. For example, nouns are allowed to participate in certain syntactic dependency roles, but not others. Put another way, dependency types are defined in part by the classes of words that they link. Due to this fact, the syntactic neighbors of a word will almost always be other words belonging to the same part of speech. However, very small word classes may represent an exception to this rule. If a language has a word class with just two members, then those words will have to look further (and beyond their own word class) to find more than two neighbors. The size of many word classes, including function words, vary greatly from language to language. What's more, some of the function words that occur in small word classes, such as conjunctions in English, are highly frequent. This may introduce strong differences in the neighborhood densities of these words across languages and interfere with the correlation that I am investigating in this study. For this reason, I restrict the analysis to nouns and verbs, which are universally large and open word classes.

3.2.1 Neighborhood density measure

To define syntactic neighborhood density, I can draw inspiration from how neighborhood densities are calculated for other features. Phonology and orthography are not particularly helpful, as the nature of these representations are very different from syntax. Since phonology and orthography are typically represented as a string of phonemes/characters, the corresponding neighborhood densities are often defined as the number of minimal pairs—words that differ from a target word by just one phoneme/character (e.g., Grainger et al., 2005; Munson and Solomon, 2004; Yates et al., 2004). This approach does not

translate to the syntactic representations because there is no such thing as a minimal pair of probability distributions.

Semantic vectors share many similarities with syntactic vectors, but this is far from the only semantic representation used in studies investigating the effects of semantic neighborhoods, and neighborhood definitions among these approaches are as diverse as the approaches themselves. These approaches can be divided into those that are object-based and language-based. Object-based approaches refer to the properties of the objects to which a word refers, and they include feature-based representations (e.g., McRae et al., 2005; Vigliocco et al., 2004). On the other hand, language-based approaches rely on the properties of the language used to refer to those objects. These include association-based representations, in which association norms are collected from participants (e.g., Nelson et al., 2004; Nelson et al., 1992) and semantic neighborhoods consist of meaningfully related words generated by those participants (e.g., Locker et al., 2003). Language-based approaches also include textual co-occurrence-based vectors (e.g., Landauer and Dumais, 1997; Lund and Burgess, 1996), which open up new possibilities for defining semantic neighborhoods. For these vectors, neighborhood densities usually refer to the average distance to a word's neighbors, but some define neighbors differently as either the words within a certain distance threshold (e.g., Danguecan and Buchanan, 2016; Macdonald, 2013) or a word's ten closest neighbors (e.g., Buchanan et al., 2001; Mirman and Magnuson, 2008; Shaoul and Westbury, 2010). It should be noted that semantic features can be predicted with high accuracy from distributional semantic vectors (Durda et al., 2009), so these different approaches may be capturing largely the same semantic information.

Drawing inspiration from this semantic precedent, I define syntactic neighborhood density as the average distance from a word to its n closest neighbors in syntactic space. Note that a high value represents a low neighborhood density, as the n closest neighbors

are far away on average. I collect such densities for three values of n: 3, 10, and 50. This allows me to explore whether effects differ significantly between small and large neighborhood sizes. Semantic and orthographic neighborhood densities are defined in the same way, using their corresponding distance metrics. For orthography, this represents a departure from the typical neighborhood definition based on minimal pairs. However, I expect it to capture very similar information, because both minimal pairs and the Levenshtein distance metric are based on the same underlying concept: single-character changes to a string. If a word has fifteen minimal pairs, the average distance to its nearest ten neighbors will be 1; in contrast, if a word has only two minimal pairs, then the average distance to its ten nearest neighbors will be substantially greater than 1 (because most of these neighbors require two or more changes). I use this new definition for consistency with the syntactic and semantic neighborhoods, and it represents an opportunity to replicate the orthographic effects found in previous research with a new neighborhood density metric.

3.2.2 Correlational analysis

To assess the relationship between syntactic neighborhood density and frequency, I first take a permutation approach. For each language, I take the Pearson correlation between these two variables (both Box-Cox transformed to approximate a normal distribution). I expect a negative correlation, because a low average distance to n neighbors represents a very dense neighborhood. As frequency goes up, I hypothesize that this neighborhood density metric will go down (i.e., increasingly dense neighborhoods).

While the Pearson correlations between these variables are insightful in themselves, it is necessary to compare them to a baseline. I create such a baseline for each language by permuting the syntactic neighborhood density variable, resulting in a random pairing

of neighborhood densities and frequencies. I performed this permutation 10,000 times for each language to produce a null distribution of correlation values against which I can compare the real correlation. In turn, I computed p-values directly from these distributions by dividing the number of simulated correlations (plus one) that are lower than the real correlation by the total number of simulated correlations (plus one). The p-value for each language represents the probability that the real correlation would have that value if it was drawn from the null distribution. This analysis was performed for each of the three values of n.

Real correlations for each language and their confidence intervals are shown in Figure 3.1, along with their significance levels. Each column in the plot represents a different neighborhood size, and within each column, correlations are shown on the x-axis. Languages are displayed on the y-axis, organized by family and subfamily.

Overall, Figure 3.1 demonstrates that—for most languages and neighborhood sizes—the correlation between syntactic neighborhood density and frequency is significantly negative. Consistent with my hypothesis, this means that more frequent words generally have more dense syntactic neighborhoods. This effect is fairly consistent across the languages, subfamilies, and families included in the study. For example, 45 out of 48 languages show a significant negative correlation for a neighborhood size of 10.

However, there are a few exceptions to the pattern in the form of languages with non-significant or even positive correlations. Some of these languages may fail to show the expected negative correlation due to a paucity of data, such as Hungarian, Welsh, and Gaelic. The wide confidence intervals for these languages are indicative of their small datasets. Hungarian, for example, was the smallest dataset in this study with only 254 qualified lemmas. (In contrast, the Czech data contains 7,590 lemmas.) Three other languages failed to display a significant negative correlation for at least one of the neighborhood sizes: Japanese, Korean, and Chinese. Chinese in particular displayed

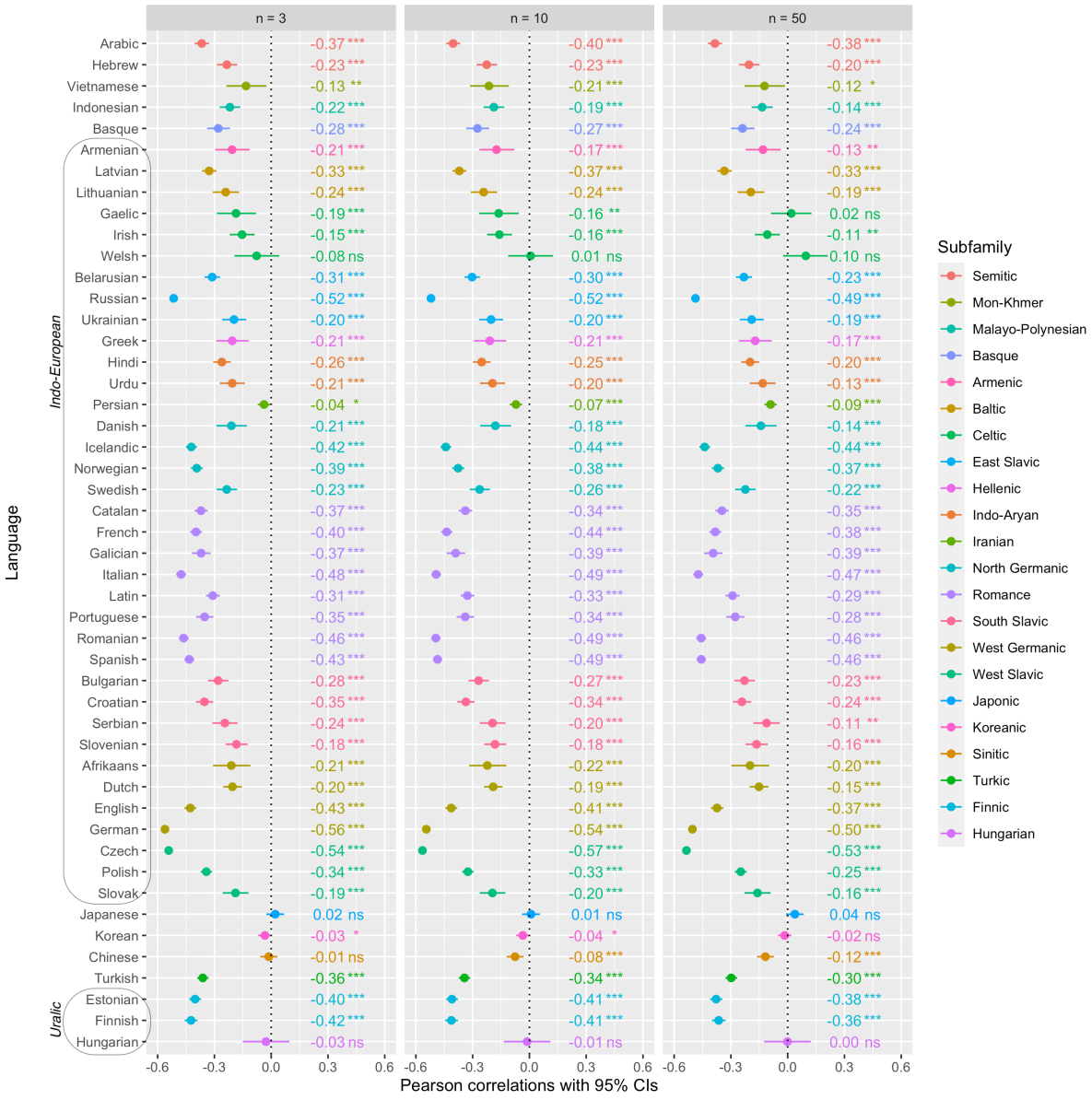

Figure 3.1: Correlations between syntactic neighborhood density and frequency in 48 languages. Correlations and their 95% confidence intervals are shown on the x-axis, while languages are displayed on the y-axis, organized by family and subfamily. Each panel corresponds to a different neighborhood size. Significance values for each language are the result of comparing the real correlation to 10,000 permutations. 45 out of 48 languages show a significant negative correlation for a neighborhood size of 10 (middle panel).

small *positive* correlations for all three neighborhood sizes. The fact that all three of these languages are spoken in East Asia raises questions about a potential areal influence, and this anomaly should be explored in future research.

I previously observed that a few languages with small datasets show non-significant correlations. It is also apparent that some of the languages with the largest datasets (indicated by tiny confidence intervals)—such as Czech, German, and Russian—exhibit some of the strongest negative correlations. This suggests a relationship between the amount of data I have for a language and the correlation that I observe. This relationship is more directly illustrated in Figure 3.2.

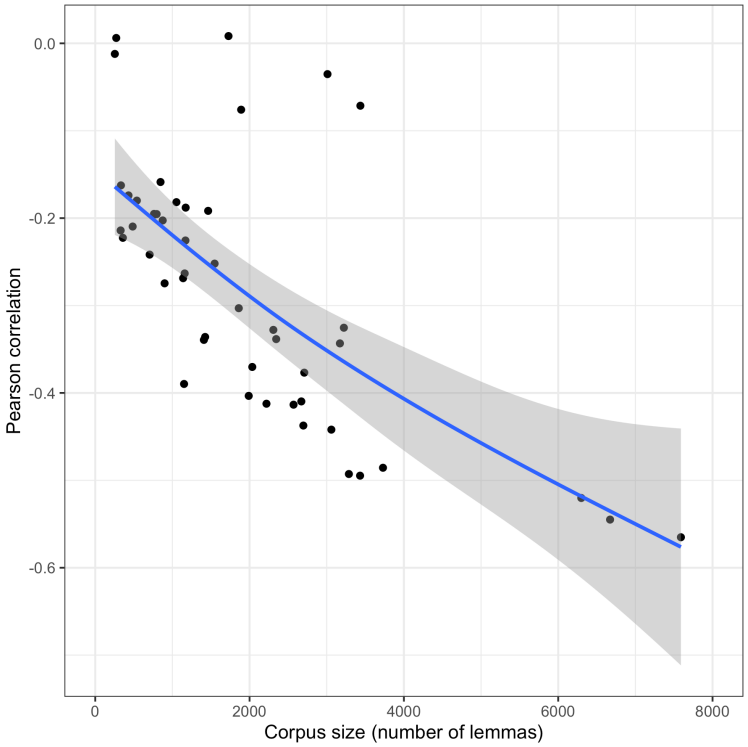

Figure 3.2: Relationship between the size of the dataset and the correlation between syntactic neighborhood density and frequency for 48 languages. Languages represented by larger datasets tend to display a lower (stronger negative) correlation.

There are a couple of possible explanations for this trend. First, the correlation of interest may be stronger at moderate word frequencies than at very high frequencies. I

only include lemmas that occur ten times in the data, but this affects languages differently depending on the size of the corpora. If the corpus for a particular language is small, then only very high frequency words (in terms of per-million frequencies) will make the cut. However, if the corpus is very large, even moderately frequent words will be attested enough times to be included. If the correlation of interest is stronger among moderately frequent words, then it will be stronger for languages with large datasets but not those with small datasets. In a similar fashion, languages with small datasets have less reliable neighborhoods. Many of the low and moderate frequency words that might participate in such neighborhoods are excluded from the analysis because they are not attested at least ten times. This problem is alleviated to some degree for languages with large corpora. Regardless of the explanation, it is encouraging that the hypothesized effect is strongest for those languages that are the most well-attested. It suggests that some of the languages which fail to show a significant negative correlation might actually exhibit one if I had more data, although future research would be needed to confirm this hypothesis. Given the existence of this relationship between dataset size and effect size, one should be cautious in generalizing from these results, particularly concerning the size of the correlation between syntactic neighborhood density and frequency.

3.2.3 Mixed-effects regression analysis

I also perform a mixed-effects regression analysis predicting frequency from syntactic, semantic, and orthographic neighborhood densities for a neighborhood size of 10. All of these variables including frequency were Box-Cox transformed and scaled within each language. I used a full random effects structure with random intercepts and slopes for languages and subfamilies. A model selection process was used to determine whether interactions and curvature should be included in the final model, and this procedure is

described in more detail in Appendix A.

The final model includes interactions between syntactic and semantic neighborhood densities and between orthographic and semantic neighborhood densities. It also includes polynomial curvature for syntax (to the 4th degree) and orthography (3rd degree), but no curvature for semantics. Predictions of this model for each of the two interactions can be seen in Figure 3.3. Neighborhood densities for the interacting variables are shown on the x and y axes, and the prediction for frequency is indicated by color. Recall that lower values for neighborhood density scales actually indicate higher densities because the numbers are based on average distance to ten nearest neighbors. For both plots, the yellows and oranges that represent high frequency predictions are found in the lower left corners. This indicates that the highest frequencies are predicted for words with the densest neighborhoods. The effect of semantic neighborhood density is consistent and strong in both plots, as colors get darker from left to right. The effect of syntactic neighborhood density can be seen in the plot on the left; the angles of contours show this effect to be strong in the center of the plot, but curved contours in certain corners indicate some unexpected patterns among extreme syntactic and semantic density values. In the plot on the right, the effect of orthographic neighborhood density is fairly consistent but weaker than those of semantics and syntax.

3.3 Discussion

In this study, I have shown that more frequent words have denser syntactic neighborhoods. This pattern is found across a large sample of languages, and for neighborhood sizes ranging from a word's three to fifty nearest neighbors. I interpret this finding as yet another example of the information-theoretic principle that the most frequent words in a language should be most optimized for efficient use (Piantadosi et al., 2011; Zipf,

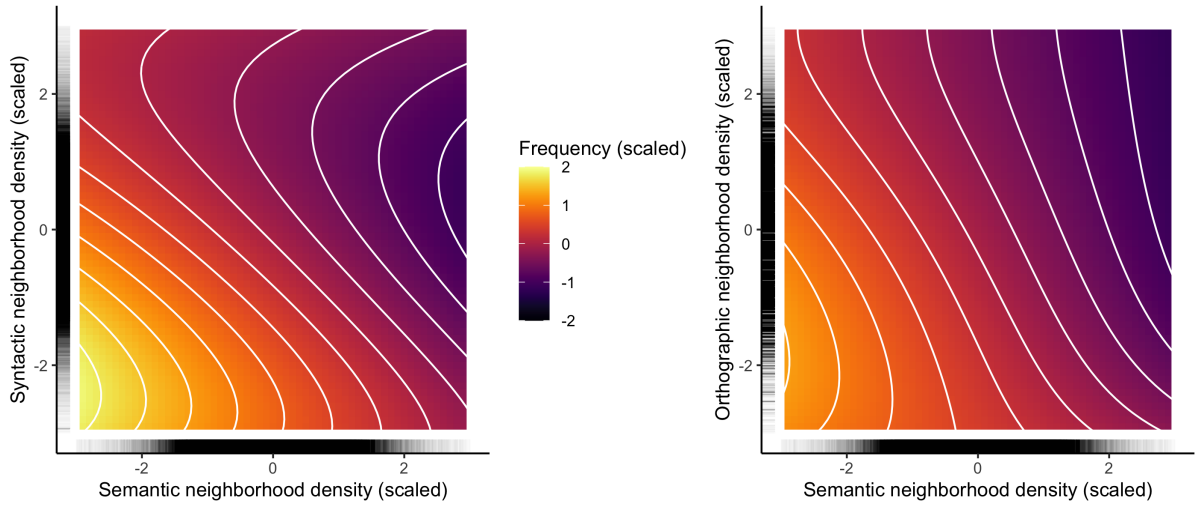

Figure 3.3: Mixed-effects model predictions for frequency showing interactions between semantic and syntactic neighborhood densities (left) and semantic and orthographic neighborhood densities (right). Densities are measured by average distance to ten nearest neighbors, so low numbers represent denser neighborhoods. Color represents predicted word frequencies. In both interactions, higher frequencies are predicted for denser neighborhoods on both axes.

1949). In this case, words are syntactically distributed within languages in such a way that the most frequent words are most syntactically optimized. The optimization here refers to the advantages to learning, memory, and production conferred by having many close neighbors.

I also reproduced previous findings that the same pattern holds for semantic and orthographic neighborhood densities. With regard to orthographic neighborhood density, I find the similar results despite using a new metric for neighborhood density. Rather than using the number of minimal pairs, I used the average distance to a word's n nearest neighbors.

Recall from earlier in this chapter that these can be seen as meta-patterns of systematicity. Many of the studies of systematicity in the lexicon—including Chapters 4 and 5 of this dissertation—focus on the existence of clustering or dispersion overall. Those represent the first level of systematicity in the lexicon. However, those patterns of clus-

tering and dispersion may be unevenly represented throughout the lexicon, and the level of clustering or dispersion may correlate with features such as frequency. This is the second level of systematicity, and it is observed in this chapter's results. I have not demonstrated a first level pattern for syntactic distributions, such as whether they are clustered over above what would be expected by chance. Yet I have shown that clustering among syntactic distributions is uneven across the lexicon, with more clustering for high frequency words.

With an understanding of first and second level patterns of systematicity, it follows that this chapter is only the first step to understanding patterns of systematicity within syntactic representations in the lexicon. I have shown how the most frequent words across languages are also the most optimized for efficient use. However, one could also ask whether the syntactic representations of words are overall clustered beyond what would be expected by chance. For example, phonological forms have been shown to cluster beyond the effects of phonotactics in real languages compared to simulated lexicons, with the comparisons based on several metrics such as average distance between wordforms, number of minimal pairs, and network measures (Dautriche et al., 2017). Of course, the nature of syntactic distributions is very different than the nature of phonological forms, so these metrics and the method for simulating baseline lexicons would need to be adapted appropriately. Similarly, Mahowald et al. (2018) demonstrated that frequent words are more orthographically well-formed than infrequent ones. One could ask an analogous question about syntactic distributions using the idea of syntactic prototypicality (e.g., Lester, 2018. Answering these additional questions about systematicity within syntactic distributions would give yield a more comprehensive understanding of how syntactic features mirror other features of the lexicon and how words are syntactically distributed for efficient use.

Given that the role of syntactic distributions in the lexicon has only recently been

investigated, there remains a need for further psycholinguistic studies to elucidate the role of this syntactic information in learning, memory, production, perception. In this study, I assume that clustering of syntactic distributions and the resulting syntactic associations between words have the same affects as those observed for other features of the lexicon. However, this remains to verified empirically. This study provides not only further motivation for such psycholinguistic studies, but also demonstrates methods for defining syntactic neighborhood density that can be used in such studies.

Chapter 4

Systematicity between syntactic distributions and other lexical features

4.1 Background

In the last chapter, I investigated a particular pattern of systematicity within syntactic distributional information. Now, I broaden my scope to investigate the relationships between syntactic distributional information and other features of words, namely semantics and phonology/orthography.

I reviewed literature in Chapter 1 that shows how the relationship between form and meaning is not as arbitrary as has been traditionally assumed (de Saussure, 1916; Hockett, 1960). Rather, there are systematic relations between form and meaning that go beyond the effects of sound symbolism and are observed across languages and language families (Dautriche et al., 2016; Monaghan et al., 2014; Shillcock et al., 2001; Tamariz, 2008). These regular correspondences appear to be functionally motivated (Dingemanse et al.,

2005), offering advantages for learning (Imai and Kita, 2014; Imai et al., 2008; Monaghan et al., 2014; Nielsen and Rendall, 2012; Nygaard et al., 2009) and memory (Kirby et al., 2015; Tamariz and Kirby, 2015). However, taken to an extreme, association between form and meaning would lead to high levels of confusability. As it is, sound-meaning correspondences offer challenges to individual word acquisition (Dautriche et al., 2015; Monaghan et al., 2011; Monaghan et al., 2012; Swingley and Aslin, 2007) and production (Dell and Reich, 1981; Goldrick and Rapp, 2002; Schwartz et al., 2006). It seems that lexicons are optimized when they exhibit limited associations between lexical features. I predict a similar pattern in the relation between syntax and these other features of the lexicon.

The relationship between syntax and semantics has been given considerable attention in linguistics research. For example, various studies have explored the relationship between the syntactic patterns and semantic characteristics of verbs (e.g., Fillmore, 2015, Levin, 1993). Construction grammar even asserts that syntactic constructions carry meaning above and beyond the lexical components of which they are composed (Goldberg, 2006). However, these approaches do not speak directly to the nature of the correlation of syntax and semantics across the lexicon. More recently, Lester et al. (2017) offer an empirical report on this relationship using fine-grained syntactic measures. Unsurprisingly, they found a significant positive correlation between syntactic and semantic distances across pairs of words in English. I seek to extend this finding to a large sample of languages.

On the other hand, the relationship between syntax and phonology has received relatively little attention. Research into this relation is limited to categorical treatments of syntax; that is, they are based on word classes rather than the fine-grained syntactic distributional vectors. Research has demonstrated statistical patterns in phonological form that correspond to grammatical distinctions such as noun vs. verb and open vs.

closed word class (Cassidy and Kelly, 1991; Kelly, 1992; Monaghan and Christiansen, 2008; Monaghan et al., 2005; Monaghan et al., 2007). These phonological cues have been shown to support category learning and generalization to novel words (Fitneva et al., 2009; Monaghan et al., 2011). However, it is unclear whether similar correspondences exist *within* word classes based on more subtle, usage-based syntactic differences.

To my knowledge, there has been no study investigating the relation between syntactic and phonological distance in the lexicon using fine-grained syntactic representations. I predict a positive correlation, just as both of these features have been shown to be correlated positively with semantics. If this is the case, it would constitute additional evidence for non-arbitrary relationships among features of the lexicon. The positive correlation would mean that each of these lexical features—phonology, semantics, and syntax—is predictable to some extent from the others. As I discussed in Chapter 1, these kinds of associations among features of the lexicon can serve to scaffold learning and reduce memory demands. The purpose of this chapter is to find out whether syntactic distributional information participates in these patterns of systematicity among features of the lexicon.

4.2 Analysis

As laid out in Chapter 2, it can be useful to pair each lemma in a language with each other lemma. This reduces lengthy strings and vectors to the set of distances representing each feature: syntax, semantics, and orthography (as a proxy for phonology). I can then compare this distance metrics across all the pairs in a language to see if they correlate with each other.

For this study, I took additional steps to clean up the paired data. First, I only include pairs of lemmas belonging to the same part of speech. For one, the syntactic de-

pendencies that make up the syntactic distributions are co-defined with parts of speech, so the distributions of lemmas from different parts of speech will essentially be mutually exclusive. Distance measures between these mutually exclusive distributions would be fairly uninformative. Second, I just discussed how previous research has already demonstrated correspondences between phonology and syntax at the level of word classes. I am instead interested in whether such correspondences exist within word classes as well.

Second, I only include pairs of lemmas of the same length. This follows the precedent of Dautriche et al. (2016), and the purpose is to minimize the comparison of morphologically related words. Since I am dealing with lemmas in this study, inflectional variants such as *dog* and *dogs* should not be a problem, but I still wish to minimize the comparison of derivational variants such as *happy* and *unhappy*.

I also subset the data to include only large open word classes (noun, verb, adjective, adverb, proper noun) and numerals (since they were fairly well-represented). Lemmas are only paired with other lemmas within the same part of speech, so I wish to remove the potential bias that can arise from having very limited pairings from small word classes.

After trimming the data in these ways, I removed languages that did not have at least 1,000 lemma pairs. This is a fairly low cut-off, so I pay special attentive to the possibility that these small sample languages may exhibit exceptional patterns due to the paucity of data. The resulting sample for this study is 48 languages, and these vary substantially with regard to the number of lemma pairs.

4.2.1 Correlational analysis

I start with a permutation approach to investigate the relationship between syntactic distance to semantic and orthographic distances. Within each language, I transform and scale each of these three distances before taking the Pearson correlation between syntactic

distance and the other two. Based on the relationship between semantics and phonology, I expect to find positive correlations between syntax and these features.

To compare these real correlations to a baseline, I permute the syntactic distance variable 10,000 times for each language, taking the correlations of interest after each permutation. The results are null distributions of correlation values for each language against which I can compare the real correlations. I computed p-values directly from these distributions by dividing the number of simulated correlations (plus one) that are higher than the real correlation by the total number of simulated correlations (plus one). The p-value for each language and correlation represents the probability that the real correlation would have that value if it was drawn from the null distribution.

Real correlations for each language and their confidence intervals are shown in Figure 4.1, along with their significance levels. The left column represents correlations between syntax and orthography, while the right column represents correlations between syntax and semantics. Correlations are displayed on the x-axis, while languages are displayed on the y-axis, organized by family and subfamily.

As expected, the correlation between syntactic and semantic distances is overwhelmingly positive. Every language in the study exhibits a highly significant, positive correlation for this relationship. In contrast, the correlation between syntactic and orthographic distances shows substantial variation across languages, with 32 languages exhibiting a positive correlation and 16 languages exhibiting a negative correlation.

4.2.2 Mixed-effects regression analysis

To further explicate the relationship between syntactic distance and orthographic and semantic distances, I also subjected the data to a mixed-effects regression analysis. I fit a model predicting syntactic distance from orthographic and semantic distances, with all

Figure 4.1: Pearson correlations between syntactic and orthographic distances (left column) and syntactic and semantic distances (right column) in 48 languages. Correlations are arranged on the x-axis, while languages are displayed on the y-axis, organized by family and subfamily. Subfamilies are also represented by different colors. The correlation between syntactic and orthographic distances is positive in two-thirds of the languages, with variable levels of significance. The correlation between syntactic and semantic distances is positive and highly significant in every language.

three of these variables Box-Cox transformed and scaled within languages. The model was equipped with a maximal random-effects structure with nested random effects for family, subfamily, and language. For languages with more than 10,000 pairs, a sample of 10,000 rows was used in fitting the regression model.

Model coefficients reveal a strong, positive effect of semantic distance on syntactic distance. The coefficient of 0.303 indicates that for 1 standard deviation increase in semantic distance, the model predicts a 0.303 standard deviation increase in syntactic distance. However, the model also indicates a small, negative effect of orthographic distance on syntactic distance. The coefficient of -0.012 indicates that for 1 standard deviation increase in orthographic distance, the model predicts a -0.012 decrease in syntactic distance. I compared this model to two others, each with one of the fixed effects removed (although maintaining the same random effects structure). Unsurprisingly, the inclusion of semantic distance is justified ($\chi^2(1) = 30.1, p < .0001^{***}$), while the inclusion of orthographic distance does not significantly improve the model ($\chi^2(1) = 0.035, p = .852$).

4.3 Discussion

The results of this study fail to support my hypothesis that syntactic distributional information would be positively correlated with both semantics and orthography (as a proxy for phonology). I found a robust positive correlation between syntactic and semantic distances, but this follows somewhat trivially from the well-known fact that distributional semantic vectors capture at least some syntactic information along with other aspects of meaning. Of particular interest in this study was the relationship between syntactic and orthographic distances. The correlational analysis revealed a general trend toward a positive correlation, with two-thirds of languages exhibiting a correlation greater than 0. However, the regression analysis revealed that this may be an artifact

of the positive correlations between syntax and semantics and between semantics and orthography. When controlling for the contribution of semantic distance, an increase in orthographic distance actually predicts a small *decrease* in syntactic distance.

The most straightforward explanation for the results concerning syntax and orthography is that syntactic distributional information does participate in any systematic relationship with orthography *beneath the level of word class*. (Recall that previous studies have shown that at least some differences in word class correspond to differences in word forms (Cassidy and Kelly, 1991; Kelly, 1992; Monaghan and Christiansen, 2008; Monaghan et al., 2005; Monaghan et al., 2007).) Put another way, these results show that the relationship between the form of a word and its use in syntactic contexts is arbitrary at this level. Juxtaposed against research showing that semantics and phonology are correlated (Dautriche et al., 2016), this finding suggests that syntax and does not pattern systematically in the lexicon in the same way as these other features, or at least that the relationship between syntax and orthography/phonology does not.

One way to reconcile this finding with the research on systematicity is to point out that syntactic distributional information *does* actually correlate with phonology, but only at the level of word classes. After all, syntactic representations are unique (compared to semantic and phonological features) in the sense that there is a clear hierarchy of distinctions. The fine-grained differences between syntactic probability vectors are nested within categorical word class distinctions. One could make the case for something similar within semantic representations (e.g., objects vs. actions), but these are far less rigid and much harder to define. Within phonological representations, perhaps shared phonological features could be used to represent more subtle measures of dissimilarity than all-or-nothing phoneme/character differences. Either way, syntactic distributional information has a unique, hierarchical internal structure that might serve to accommodate functional pressures in different ways. I've discussed how the correlation between form and meaning

is observable across languages, but that it is also tempered by pressures of dispersion such that this relationship remains largely arbitrary. After all, systematic associations between form and meaning taken to an extreme would lead to highly confusable words across the lexicon. The same balance between correlation and arbitrariness in the relationship between syntax and orthography/phonology may be mediated by the word class hierarchy. Correlations between word classes and word forms cue category learning, while arbitrary relations within word classes facilitate perceptual distinctiveness.

What's more, this narrative calls for a re-evaluation of other relationships in the lexicon such as the one between form and meaning. It's possible that an analogous pattern exists in this relationship, with observed associations between form and meaning driven by macro-level semantic differences while more subtle semantic differences do not correspond to phonological differences. There's some evidence for the macro-level form-meaning correspondences, such as phonological correlates to the semantic distinction between concrete and abstract words (Reilly et al., 2012). However, it's unclear whether these associations disappear in finer-grained semantic distinctions. It might be possible to determine this by reanalyzing the data used in studies on the association between form and meaning.

Another possible explanation for the findings in this chapter is that syntax is different from other features of the lexicon. Syntax represents the context in which words are encountered by language users, and for listeners in particular, this is the context in which words must be disambiguated. For this reason, pressures toward distinctiveness may be particularly strong for syntactic representations. These pressures would work against the sort of association that has been observed between form and meaning (and that I predicted for the relationships between syntax and these other features). For words that tend to occur in the same syntactic environments, being additionally similar in phonological form could result in confusability. In this way, there may be functional

reasons to avoid a strong correlation between syntax and form. In a vaccuum, this sort of pressure would not just predict an arbitrary relationship, but would predict a negative correlation. To minimize confusability, as words become more similar in one dimension, they should become more distinct in the other. However, the results show almost no correlation between syntax and orthography. It is plausible that the pressures of association that lead to the correlations I see elsewhere are cancelled out by pressures of distinctiveness unique to syntax. The result—at least for the relationship between syntax and wordform—would be the absence of any strong correlation in one direction or another. Of course, this explanation cannot be support by this current approach, so it remains speculative.

Chapter 5

The role of syntactic distributions in grammatical gender assignment

5.1 Background

Grammatical gender has often been derided as an apparently arbitrary and unnecessary feature of language, perhaps most famously by Mark Twain in 'The Awful German Language' (1880): "In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl..." In languages with grammatical gender, nouns belong to two or more classes based on the agreement patterns they trigger in associated words. However, languages vary widely in their rules for assigning nouns to different genders (Corbett, 1991) and these rules are often broken by conspicuous exceptions such as the ones highlighted by Twain.

Perhaps because of this reputation, linguists have long sought to understand what advantages grammatical gender might offer to language users. After all, how could such systems arise and persist in so many of the world's languages if they served no purpose?

For one, gender has been credited for linking temporally separated elements in discourse in languages with more flexible word orders such as Latin (Dye et al., 2017). In a similar way, gender is thought to aid reference tracking in discourse by linking gendered anaphoric pronouns to the correct antecedent (Heath, 1975; Zubin and Köpcke, 1986). However, these explanations do not apply to all languages or even all cases of ambiguity (Dye et al., 2017: p. 3).

Alternatively, accounts rooted in information theory continue to offer promising ideas concerning the functional advantages of gender. There are a number of psycholinguistic studies that suggest gendered articles can guide lexical prediction (Arnon and Ramscar, 2012; Bates et al., 1996; Grosjean et al., 1994; Schriefers, 1993; Van Berkum et al., 2005; Wicha et al., 2004). Some of these studies speak to the finer cognitive mechanisms underlying the boost to prediction such as the roles of facilitation and inhibition, but the general logic is straightforward: if the gender of a noun is revealed in a preceding element, the list of candidates that might fill that noun slot is reduced significantly. A recent corpus study of German provides empirical support for this theory, showing that gender marking on German articles serves to reduce the entropy (uncertainty) of upcoming nouns (Dye et al., 2017; Futrell, 2010). Adjectives may serve the same purpose in English, a language without gender (Dye et al., 2018). These findings are consistent with information-theoretic predictions and research showing that speakers modulate speech in various ways to reduce excessive peaks and troughs in information density (Jaeger, 2010; Levy and Jaeger, 2007; Levy, 2008a).

If reducing the possible set of candidate words is a general strategy for guiding lexical prediction, then a more direct strategy would be to target those candidates that are the most likely alternatives to the intended word. Put another way, the most efficient way to lower the uncertainty of an upcoming noun is to eliminate its strongest competitors. So, what kinds of nouns compete most strongly in lexical prediction?

One proposal suggests that it is semantically similar words that compete most strongly in this way. On the one hand, semantically similar words have been shown to cluster within genders across languages (Corbett, 1991). This is even true of inanimate nouns that fall outside of the semantically transparent semantic core of animate nouns (Williams et al., 2019). On the other hand, exceptions abound, and these exceptions have been cited as evidence for the discriminatory role of gender. In a lengthy discussion of the complex relationship between semantics and gender assignment, Dye et al. (2017) argue that the German gender system combines semantic clustering and semantic dispersal. If semantically similar nouns are largely clustered within genders, the assignment of some high frequency nouns to different genders would provide the most efficient reduction in entropy when gender tips its hand. The authors cite German words for drinks as an example. The words for beer (*Bier*) and water (*Wasser*) are neuter, while most other words for drinks in German are masculine (e.g., *Wein* 'wine', *Kaffee* 'coffee', *Tee* 'tea', etc.). Once the gender of a drink is revealed, listeners can safely eliminate either the two most predictable candidates or all the rest. Compare this scenario to one in which two low frequency drinks are the gender assignment exceptions; unless one of those two low frequency drinks are intended—which would be unlikely based on their low frequency–the reduction in entropy that comes with knowing the gender is minimal, only eliminating two candidates that were already improbable. In this way, semantic clustering of low frequency words and semantic dispersal of certain high frequency words can benefit discrimination. The authors found evidence for this kind of pattern across the German lexicon: high-frequency nouns tend to be distributed across genders in German, while low-frequency nouns tend to be clustered within the same gender.

Alternatively, one could argue it is phonologically similar words that compete most strongly in lexical prediction because they are potentially confusable, particularly from the perspective of noisy channel models (e.g., Levy, 2008b). However, it does not seem

to be the case that gender discriminates such words. It is well known that gender is often marked phonologically on nouns (Corbett, 1991). The phonological rules for gender assignment vary widely from language to language, but within a given language, the nouns that share a particular diagnostic phonological pattern are overwhelmingly assigned to the same gender. To cite a familiar example, nouns in Spanish ending in *-o* are almost always masculine, while those ending in *-a* are almost always feminine. Therefore, it does not appear that grammatical gender disambiguates phonologically similar nouns.

In this chapter, I argue that these previous accounts are actually missing a fundamental piece of the puzzle. It may not be very useful to ask whether gender helps to disambiguate semantically or phonologically similar words if one does not also control for syntax. I propose syntax as the locus of disambiguation because it represents the crucial context within which words must be discriminated. Nouns that tend to occur in the same syntactic contexts will compete for activation more than nouns that tend to occur in different syntactic contexts. Thus, if a primary function of grammatical gender is to guide lexical prediction, I hypothesize that nouns occurring in similar syntactic contexts should be less likely to share a grammatical gender. In this way, some of the strongest competitors of a target noun would be eliminated at the first indication of the word's gender. Like some of the studies reviewed above, such a pattern would be probabilistic in nature, operating statistically across the lexicon, and yet it would constitute further evidence for functionally motivated structure underlying seemingly arbitrary grammatical gender systems.

Using the UDT data and methods introduced in Chapter 2, I test the prediction that—across a large sample of languages—nouns will be assigned to genders such that gender supports the disambiguation of syntactically similar words. Put simply, syntactically similar words should be less likely to share gender than syntactically dissimilar words. Based on the literature reviewed above, I expect the opposite pattern for seman-

tically and phonologically similar words.

5.2 Analysis

5.2.1 Correlational analysis

To assess the relationship between syntactic distance and gender sameness in pairs of lemmas, I first take a permutation approach. One straightforward way to perform such an analysis would be to permute the syntactic distances for a language in the paired lemma data and then calculate the correlation of this permuted variable with gender sameness. Performing this permutation many times would produce a null distribution of correlation values against which I could compare the real correlation.

However, this approach is complicated by systematic relationships between each of these variables and secondary variables in the data: semantic and orthographic distances. The relations of syntactic distributions to both form and meaning are investigated and discussed at length in Chapter 4. The top panels of Figure 5.1 show the Pearson correlations between syntactic distance and both semantic and orthographic distances in the languages of this gender study: correlation values are shown on the x-axis, and the number of languages that display those correlations is shown on the y-axis. Syntactic and semantic distances are positively correlated in every one of these languages, while syntactic and orthographic distances are positively correlated in over two-thirds of the languages. These correlations show that, in general, syntactically similar words are also more likely to be semantically and orthographically similar.

Additionally, I know from the literature that both phonology (and its proxy orthography) and semantics are implicated in gender assignment cross-linguistically. Shared phonological patterns can indicate shared membership in a particular gender (Corbett,

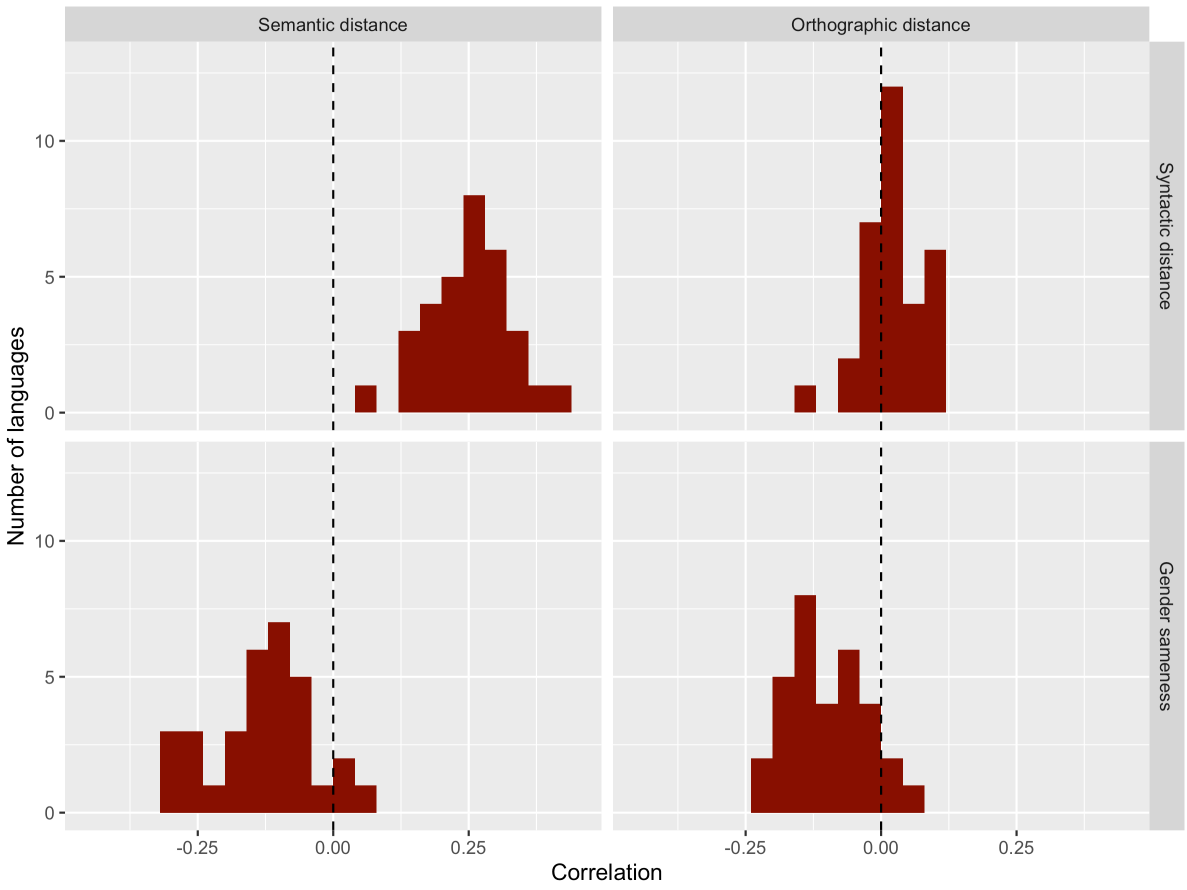

Figure 5.1: Correlations between the variables of interest (syntactic distance and gender sameness) and secondary variables (semantic distance and orthographic distance) among the 32 languages of this study. Semantic and syntactic distances are correlated positively in every language, while orthographic and syntactic distances are correlated positively in more than two-thirds of the languages. Both semantic and orthographic distances are correlated negatively with gender sameness in over 90% of the languages.

1991). Likewise, semantically similar words have been shown to be more likely to share a gender across the lexicon (Williams et al., 2019). These observations from the literature are borne out in my data, as illustrated in the bottom panels of Figure 3. Both semantic and orthographic distances are correlated negatively with gender sameness in over 90 percent of the languages. These negative correlations mean that as nouns become more semantically or orthographically distant, they are less likely to share a gender.

These patterns of systematicity can help one predict the relationship that would be expected by chance between syntactic distance and gender sameness. If syntactic distance is correlated positively with both semantic and orthographic distances, and in turn these variables are both correlated negatively with gender sameness, then—all else being equal—one should also expect syntactic distance to have a negative correlation with gender sameness. My goal is to adjust the null distribution of the correlation between syntactic distance and gender sameness to account for this systematicity elsewhere in the data. To accomplish this, I develop a variation on correlational analysis, an algorithm that I will refer to as controlled permutation.

Just like a typical permutation analysis, controlled permutation begins with a random permutation of the variable of interest—in this case the syntactic distances in the data of a particular language. However, before calculating the correlation of interest, the algorithm works incrementally to restore the known correlations between the permuted variable and the secondary variables up to a user-specified degree of tolerance (precision). Two rows of the data are chosen at random, and the algorithm evaluates whether swapping the syntactic distances of those rows would push the correlations in the desired direction. If so, the switch is made; if not, no change is made and a new pair of rows are chosen at random. These swaps continue until the original correlations with the secondary variables are restored within the desired tolerance level. In the data, the algorithm is complete when the original correlations between syntactic and both semantic and orthographic

distances are restored with a tolerance of ±0.001. At that point, the correlation between syntactic distance and gender sameness is calculated. As in other permutation analyses, this process is repeated many times to create a null distribution; specifically, I conducted 10,000 controlled permutations on each language. Each simulation was performed on a random sample of 10,000 rows of the data for that language, and a different sample was obtained for each simulation. I obtained one-sided p-values directly from the distribution, calculated as the number of correlations plus one that were greater than or equal to the true correlation, divided by the total number of correlations plus one (totaling 10,001; Davison and Hinkley, 1997; North et al., 2002: 439).

The results of the controlled permutation analysis can be found in Figure 5.2. I found that the correlation value between syntactic distance and probability of gender sameness is significantly greater than expected by chance in 25 out of 32 languages. Since the syntactic variable is a distance measure (rather than a measure of similarity), a greater-than-chance correlation means that syntactically similar nouns are less likely to share a gender than expected. Of the remaining 7 languages, only 1 shows a correlation significantly lower than expected by chance. This outlier is Latin, whose status as an extinct language (Eberhard et al., 2021) calls into question the nature of its corpus; source materials for Latin are less likely to reflect naturalistic data, and this offers a plausible explanation for its aberrant place among these results.

It is important to note that a correlation significantly greater than chance does not necessarily mean a positive correlation. In fact, many of the significant correlations in Figure 4 are below zero. Put another way, it is not always true that syntactically similar nouns are less likely to share a gender than syntactically dissimilar nouns. For some languages the opposite is true, even if only slightly. However, when the correlation is significantly greater than chance, then one can say that syntactically similar nouns are assigned to the same gender less often than expected, all else being equal.

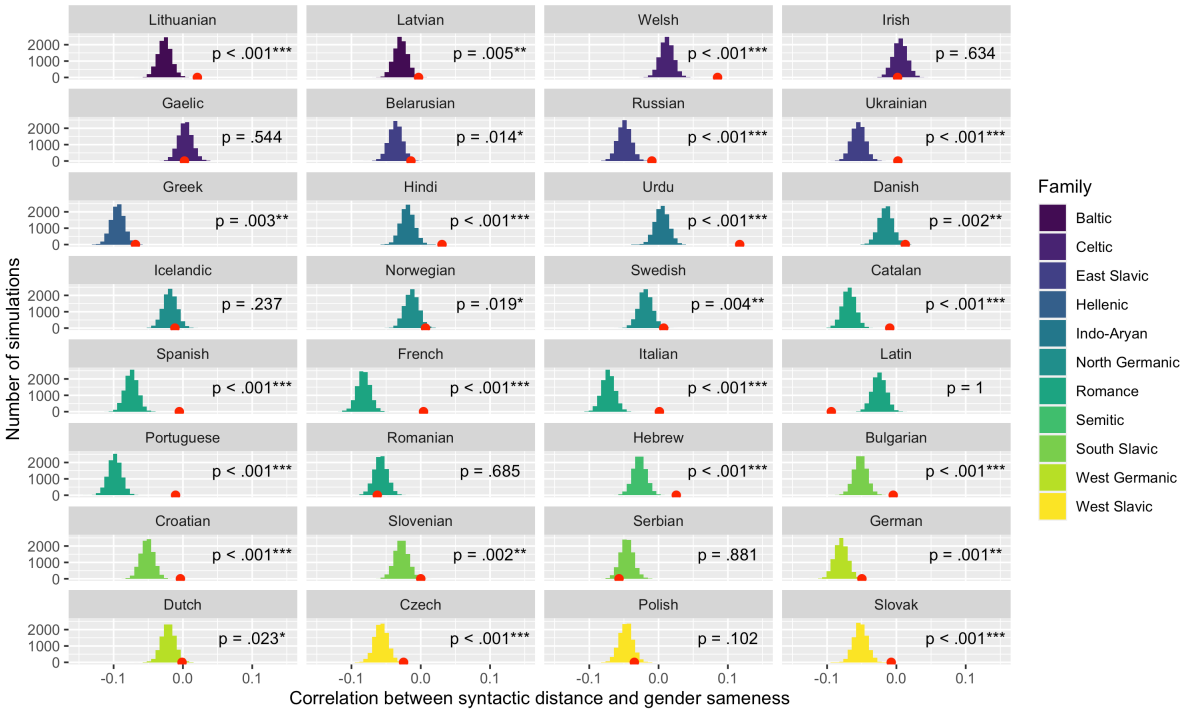

Figure 5.2: Controlled permutation analysis of the correlation between syntactic distance and gender sameness in lemma pairs of 32 languages. The red dots represent the true correlations observed in the data, while the histograms represent simulated correlations. Language families are represented by different colors. For 25 of 32 languages, the real correlation value is significantly greater than expected by chance.

5.2.2 Mixed-effects regression analysis

In addition to the permutation analysis, I fit a mixed-effects generalized linear regression model predicting gender sameness (*no* vs. *yes*) from orthographic, semantic, and syntactic distances and number of genders (a factor distinguishing languages with two vs. three genders). The regression was fit on randomly sampled parts of the data consisting of 10,000 pairs for each language, and each of the numeric predictors were Box-Cox normalized (Fox and Weisberg, 2019: Section 3.4.2) and scaled within each language. I included random intercepts for each language and language family, as well as random slopes for each of the fixed effects for both of these grouping levels. This random effects structure allows the influence of each predictor on the dependent variable to vary across languages and families. In other words, the model can reveal which effects are language- or family-specific, and which ones persist cross-linguistically. This general modeling approach follows recent studies on lexical phenomena using similarly large language samples (e.g., Dautriche et al., 2016; Mahowald et al., 2018).

Model coefficients reveal the following effects for each fixed-effect predictor on the dependent variable:

- An increase in orthographic distance predicts a decrease in probability of gender sameness
- An increase in semantic distance predicts a decrease in probability of gender sameness
- An increase in syntactic distance predicts an increase in probability of gender sameness
- The probability of gender sameness is lower in three-gender languages than it is in two-gender languages

In other words, semantically and orthographically similar nouns are more likely to share a gender, but syntactically similar nouns are less likely to share a gender. These effects are illustrated in Figure 5.3. Inspection of the random effects indicates that the overall effect of syntactic distance is not attributable to just one or a few language families. Consistent with the correlational analysis, there is some language-specific variation in this effect, but language families do not vary substantially. Likelihood ratio tests comparing this model to four additional ones—each with one of the fixed effects removed (but no change to random effects)—indicate that the full model explains the data better than ones without a fixed effect for semantic distance ($\chi^2(1) = 12.8, p < .001^{***}$), orthographic distance ($\chi^2(1) = 11, p < .001^{***}$), syntactic distance ($\chi^2(1) = 17.5, p < .0001^{***}$), and number of genders ($\chi^2(1) = 19.3, p < .0001^{***}$). Additional regression details can be found in Appendix A.

The effect of number of genders on gender sameness follows logically from the principle that—all else being equal—a greater number of classes means it will be less likely that two randomly chosen elements belong to the same class. However, I also want to consider the possibility that the overall effect of syntactic distance on gender sameness varies based on the number of genders in a language. Hypothetically, this effect could be strong for two-gender languages but disappear for three-gender languages, or vice versa. To test whether this is the case, I fit a model with an interaction between syntactic distance and number of genders as an additional fixed effect, along with corresponding random slopes for language and family. A likelihood ratio test comparing this new model to one without the interaction (but no change to random effects) indicates that this interaction does not significantly improve the model ($\chi^2(1) = 0.025, p < .874$). This means that the effect of syntactic distance on gender sameness does not vary significantly between two- and three-gender languages.

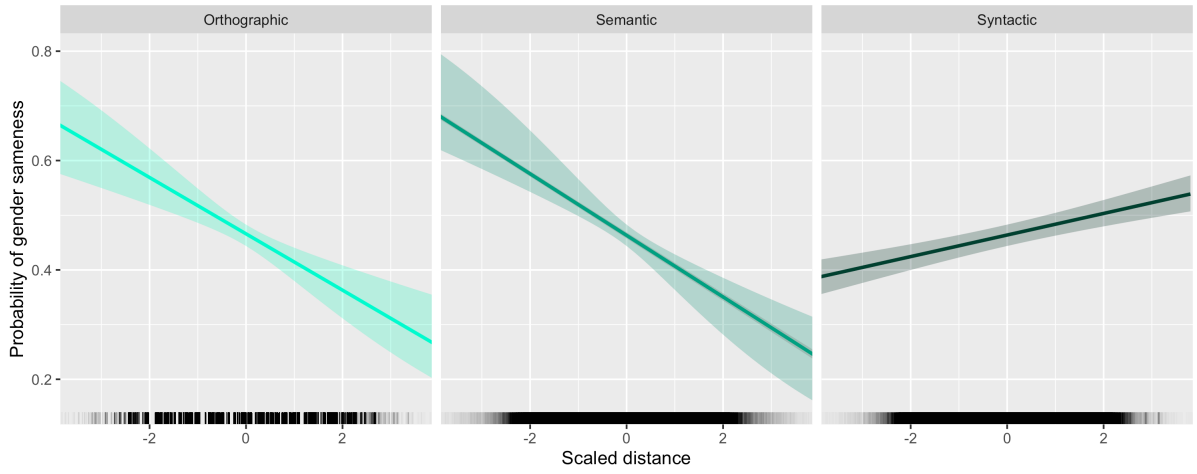

Figure 5.3: Fixed effect plots showing the influence of orthographic, semantic, and syntactic distances on the probability of gender sameness between pairs of nouns. The first two panels show that as orthographic and semantic distances increase, probability of gender sameness decreases. The third panel shows the opposite pattern: as syntactic distance increases, probability of gender sameness also increases.

5.2.3 Discussion

These results show that, cross-linguistically, syntactically similar nouns are assigned to the same gender less often than syntactically distant nouns. This relationship between syntactic distance and gender sameness is exactly the opposite of the one I find for semantics and orthography. This pattern persists across a large sample of languages, and it is not driven by just one or a few languages or language families.

I interpret this finding concerning syntax as a reflection of information-theoretic pressures on language. By definition, syntactically similar words tend to occur in the same syntactic contexts, and therefore they compete against each other for activation in these contexts. A grammatical mechanism which disambiguates such words would be advantageous to language users, curbing confusability and facilitating more accurate lexical comprehension. It appears that grammatical gender serves this very role. Those syntactically similar words that compete most strongly with each other tend to be distributed across genders rather than within them. Grammatical gender has been shown to guide

lexical prediction by reducing the set of candidate nouns that can occur following a gender-revealing preceding element, and I have shown that this candidate reduction process eliminates some of the strongest syntactic competitors of the target word. The apparently arbitrary system by which nouns are assigned to genders across languages may instead be a design feature of those languages.

In Chapter 1, I reviewed the ways in which patterns of systematicity in the lexicon reflect both pressures of association and dispersion. The results of this study on grammatical gender assignment are consistent with this view of the lexicon. Grammatical gender systems exhibit the same patterns of association and dispersion seen elsewhere in the lexicon. The well-documented rules relating semantic and phonological features to gender assignment reflect pressures of association. Grouping semantically and phonologically similar nouns together within genders likely serves to scaffold learning and reduce memory demands. For example, knowledge of these associations would allow a language learner to correctly infer the gender of a noun more often than in an arbitrary system. What's more, studies showing that speakers often know the gender of the incipient word in tip-of-the-tongue situations suggest that the association with gender may facilitate access to lexical items (Vigliocco et al., 1997). Yet, as I discussed earlier, the story may be more complicated with semantics. This largely taxonomic system may be interlaced with strategic exceptions in gender assignment in the form of high frequency words that aid discrimination (Dye et al., 2017).

The primary contribution of this chapter has been to further demonstrate how gender systems reflect information-theoretic pressures. Nouns are distributed among genders in such a way as to minimize confusability between targets and their syntactically similar competitors. These advantages are most salient for the hearer who is tasked with discriminating the intended message from possible alternatives. Thus, grammatical gender systems serve as a microcosm of the lexicon as whole, shaped by competing forces.

Perhaps the genius of this functional negotiation is in the way opposing pressures are accommodated in different ways and in different dimensions of the lexicon.

Chapter 6

Conclusions

In this dissertation, I have set out to investigate the ways in which syntactic distributional information patterns in the lexicon. This research question was borne out of the intersection of two recent threads in linguistic research. On the one hand, a growing number of studies are emerging that describe patterns of systematicity in the lexicon (Dingemanse et al., 2005). These statistical patterns within and among lexical features seem to be driven by functional pressures on language acquisition and use, supporting the idea that the lexicon is structured for language efficiency. On the other hand, Lester and colleagues have demonstrated that our knowledge of words includes rich, probabilistic information about the syntactic tendencies of those words (Lester, 2018; Lester et al., 2017; Lester and Moscoso del Prado Martín, 2016). This finding situates lexical items in a multidimensional syntactic space, and it puts syntax in the same league as other lexical features with regard to complexity.

Where these ideas meet, I find a gap in the literature. To my knowledge, no previous research has sought to understand the role of syntactic distributional in these patterns of systematicity within the lexicon. Each study in the dissertation addresses a different aspect of this question: How does syntactic distributional information pattern internally

(Chapter 3), with regard to other features of the lexicon (Chapter 4), and with regard to other lexically-related grammatical phenomena (Chapter 5)? On the whole, I find convincing evidence that syntactic distributional information is patterned in non-arbitrary ways. These patterns seem to reflect the roles of syntax as both a feature of the lexicon and the context within which words must be disambiguated.

In Chapter 3, I found that more frequent words have denser syntactic neighborhoods, in addition to denser orthographic and semantic neighborhoods. This pattern of systematicity reflects pressures for clustering/association, but at a meta level; the observation is that the amount of clustering is correlated to frequency. This relationship was previously demonstrated for phonological neighborhood density and frequency (Mahowald et al., 2018), and the authors interpreted the finding as evidence that the most frequent words are also most optimized for communication. This is because, for words, having many near phonological neighbors has been shown to facilitate learning, memory, and production. I propose the same interpretation for these results: words are syntactically distributed for efficient use. More broadly, the results in Chapter 3 show how syntactic distributions pattern in the same way as other lexical features. This represents a new kind of evidence in support of the psycholinguistic studies demonstrating the existence of rich syntactic information in the lexicon.

Chapter 4 turned the focus from within syntactic distributional information to its relationship with other features of the lexicon. While syntax and semantics enjoy a strong positive correlation for rather trivial reasons, I did not find convincing evidence for a correlation between syntactic and orthographic distances below the level of word class. However, previous research has shown that such a correlation exists across word classes, as word class categories enjoy phonological correlates. While this result fails to live up to my predictions, it suggests new ideas for how functional pressures may be negotiated at hierarchical levels within a particular lexical feature. I propose an explanation in which

association pressures at the level of word classes are balanced by pressures toward arbitrariness within word classes. The same concept can be explored in fresh examinations of the relationship between form and meaning. Another possible explanation stems from the unique role of syntax as a locus of disambiguation. Pressures toward association may be cancelled out by pressures toward dispersion that are particular to syntax, an idea that gets more support in Chapter 5.

In Chapter 5, I brought syntactic distributional information to bear on the problem of grammatical gender assignment. Gender systems are famously riddled with seemingly arbitrary gender assignments, which has led to decades of linguistic research seeking to discover underlying reason and function in these systems. Following promising research on gender and lexical prediction, I demonstrated that gender assignment may serve to disambiguate syntactically similar words. While semantically and phonologically similar words are more likely to be grouped within genders, syntactically similar words are more likely to be placed in different genders. This finding highlights the unique role of syntax as a locus of disambiguation. It also illustrates how the overall distribution of nouns among genders reflects both pressures of association (semantic and phonological clustering) and dispersion (syntactic distinctiveness). In this way, grammatical gender systems are a microcosm of the functional negotiation within the lexicon.

Taken as a whole, the studies within this dissertation paint a novel picture of the role of syntax within the architecture of the lexicon. There is evidence that syntactic representations pattern similarly to semantic and phonological representations in some ways. This finding is consistent with the idea that the lexicon contains fine-grained information about the syntactic properties of words, just as it contains fine-grained information about word forms and meanings. Syntax is a feature of words just like these other features, and so it is not surprising that it would be subject to the same functional pressures within the lexicon. Just as words are organized in particular ways with regard to their forms

and meanings, they are similarly organized with regard to their syntactic distributions. At the same time, there is evidence that syntactic representations have a unique role in the lexicon. Although syntactic distributions are properties of words, they also represent the context within which those words must be disambiguated from other words. In this way, syntactic representations appear to be subject to pressures of dispersion in contexts where other lexical features are not, such as in grammatical gender systems. Overall, these studies make it clear that the syntactic distributions of words are not patterned arbitrarily within the lexicon.

Thus, the emerging picture is one of functional negotiation, and examples of this balance of pressures have been cited throughout the dissertation. Traditional features of the lexicon show signs of clustering and association among them, offering advantages to learning, memory, and production; yet these patterns are limited, leaving plenty of room for distinctiveness to aid perception. In a similar way, patterns of clustering within features are modulated by frequency, with the densest clusters and their communicative benefits serving the most frequently used words. In grammatical gender systems, phonologically and semantically similar words cluster within genders while syntactically similar words are dispersed across them to aid disambiguation. Even in the limited scope of this dissertation, I find evidence that syntactic distributions are subject to the same functional tug-of-war, exhibiting both clustering and dispersion in different domains. The balance of such design features within the lexicon lend support to a growing body of literature attesting to the ways in which language structure is evolved for efficient use (Christiansen and Chater, 2008; Gibson et al., 2019).

6.1 Limitations and future directions

This dissertation represents the first attempt to investigate whether fine-grained syntactic representations participate in patterns of systematicity within the lexicon. As such, there is plenty of room for further exploration and improvement of data and methods.

First, I reiterate the questions that have gone unanswered in this dissertation. For example, in Chapter 3 I investigated one particular pattern of systematicity within syntactic distributions. However, I pointed out several other patterns that could also be investigated. For example, it is still an open question as to whether syntactic distributions cluster beyond what would be expected by chance. Similar studies have been conducted on phonological forms, and those methods could be easily adapted for an investigation into syntactic clustering.

Beyond these immediate questions, this dissertation points to the dearth of research into syntax as a rich, probabilistic feature of words. There is an extensive literature on semantic, phonological, and orthographic features in the lexicon. This includes a wide range of psycholinguistic and corpus-based studies. Yet there are very few of either of these for syntactic distributions. A great deal of research on syntax is needed to fill gaps and replace inferences with empirical results. This dissertation has uncovered evidence in corpora that syntactic distributional information is structured in particular ways, but one must infer functional motivations for such structures because much of the psycholinguistic groundwork has yet to be laid. For example, I assume that dense syntactic neighborhoods confer the same advantages to language acquisition and use as dense phonological neighborhoods, but experiments are needed to verify this empirically. As linguistic theories are forced to acknowledge and accommodate the existence of rich syntactic information in the lexicon, it is my hope that more linguists are motivated to fill these gaps in the literature.

More broadly, the chapter on grammatical gender shows how attention to syntactic distributional information can bring insights to other grammatical phenomena. The previous study on grammatical case systems is yet another example (Lester et al., 2018) This direction is a bit more open-ended, as it will require creativity to determine the ways in which syntactic distributional information might relate to seemingly disparate grammatical phenomena. The question can also be reversed, instead asking how various grammatical phenomena might influence the syntactic distributions of words. This opens up additional possibilities, potentially shedding light on how syntactic distributions reflect other linguistic structures. For example, it would be interesting to look at how syntactic distributions differ across nouns in different case roles and different morpho-syntactic alignment systems. One might find that subjects tend to take different dependents than objects. If this is the case, then one could ask if intransitive arguments differ in their syntactic distributions depending on the alignment of the language. Do nominative intransitive arguments pattern more like transitive subjects, while absolutive intransitive arguments pattern more like transitive objects? A finding like this would indicate that syntactic usage patterns of words are influenced by the structures imposed by a particular language.

6.1.1 Data and methods

One of the biggest challenges to the studies of this dissertation is availability of data and annotations. The Universal Dependencies Treebanks (de Marneffe et al., 2021) is a fantastic resource for syntactic dependencies in many languages, but it remains to be improved in various ways.

The amount of available data is a key concern for studies of the lexicon. The best investigations of phonology and semantics utilize very large datasets (e.g., Dautriche et

al., 2016). This is necessary to explore a wide range of words, and to obtain quality representations of low-frequency words. The larger the corpora, the more lexical items will be included and the more reliable the representations of those lexical items will be. For example, current state-of-the-art semantic word embeddings used in cutting-edge technologies are built on massive corpora (e.g., Devlin et al., 2019). The syntactic representations used in this dissertation are limited, extracted from the modest and even small corpora of the UDT. Much larger corpora than the UDT are available for every language in this study, but they do not offer the syntactic annotations necessary for extracting syntactic distributional information. Similarly, the measures of frequency and dispersion used in these studies suffer from the same problem.

Due to the corpus size problem, I used pre-trained semantic vectors from fastText (Bojanowski et al., 2016). Yet this process comes with its own challenges. For one, we are matching syntactic and semantic representations from different sources, which may introduce incongruities that effect the analyses. Matching wordform-based semantic vectors to UDT lemmas likely introduces additional noise. I use weighted averages to accomplish this, but the weights are based on the imperfect frequencies in the UDT corpora.

Additionally, I have used orthography as a proxy for phonology throughout this study. While other studies have done the same, ideally these studies would be reproduced with phonological codes when they become available for more languages. For now, the results could be replicated on a select few languages for which phonological codes are already available.

The approach I took in this dissertation was distinctly cross-linguistic, making sacrifices in depth and language-particular detail to obtain more widely generalizable results. However, one way to overcome many of the data challenges discussed above would be to focus on a particular language for which these resources are readily available. A good

candidate would be a language like German, for which the sDeWac super-corpus of nearly a billion words is already available (Faaß and Eckart, 2013). A corpus this large could be used to train semantic word embeddings, and natural language processing tools such as the Stanford Dependency parser (Rafferty and Manning, 2008) could be applied to annotate it for lemmas and syntactic dependencies. Essentially, the size of the corpus would produce fairly reliable frequencies and representations, and it could serve as the sole source of data (rather than joining data from disparate sources). In addition, a corpus of this size would allow one to test the hypotheses of this dissertation on words of much lower frequencies where functional pressures may play out differently.

On the other hand, this cross-linguistic approach largely falls short of ideals in geographic and genealogical representation among typological studies of language (Bickel, 2008; Dryer, 1989; Miestamo et al., 2016; Rijkhoff and Bakker, 1998, inter alia). As I have mentioned, this dissertation is subject to the limitations of the data, and I am limited to languages for which extensive syntactic dependency annotations are available. Ten language families and extensive coverage within the Indo-European language family are a decent starting point, but much more can be said about what language families and regions are not represented in my sample. Indigenous languages of the Americas, Sub-Saharan Africa, and Australia are not represented, despite the incredible linguistic diversity represented within and across these regions. Throughout this dissertation, I attempt to include as many languages as possible so that I can responsibly generalize the findings beyond a single language or family. However, I acknowledge the work that must be done going forward to document languages and develop language resources that could yield an inclusive and comprehensive understanding of these phenomena in the languages of the world. As such, all the results and conclusions presented in this dissertation should be interpreted with that caveat in mind, and with the goal of replicating these findings more widely in the future.

6.1.2 Diachrony

Finally, this dissertation has taken a largely synchronic approach to understanding systematicity in the lexicon. I have only briefly mentioned the related diachronic question: how do such patterns enter and persist in the lexicon over time? Some promising research programs have begun to address this question for other features of the lexicon (see Dingemanse et al., 2005 for review). These programs arise from the understanding that words are cultural items that only persist in a language if they are efficient for communication and able to be learned (Chater and Christiansen, 2010; Enfield, 2014, 2015; Zipf, 1935). As such, computational modeling and iterated language learning experiments have been employed to explore how language structures are shaped by communication between language users and transmission to new generations of language users (Kirby and Hurford, 2002; Kirby et al., 2015; Kirby et al., 2008; Smith et al., 2003). This research has demonstrated how systematicity can arise through repeated cultural transmission in an initially arbitrary language (e.g., Silvey et al., 2015; Winters et al., 2015). Similar methods could be used to address the evolution of syntactic distributions of words. A key question in this endeavor is whether macro-processes play out largely through lexical replacement or through lexical change. In other words, can these diachronic patterns be explained simply by the replacement of words that do not conform to functionally-advantageous patterns, or do the syntactic distributions of particular words change over time (i.e., people gradually change the ways in which they use words) in response to functional pressures? With regard to syntactic distributional information, these questions remained unexplored; yet they will need to be addressed if one wants a comprehensive understanding of the architecture of the lexicon and its motivations.

Appendix A

Additional regression model details

A.1 Chapter 3 regression details

I performed a mixed-effects regression analysis predicting frequency from syntactic, semantic, and orthographic neighborhood densities for a neighborhood size of 10. All of these variables including frequency were Box-Cox transformed and scaled within each language. I used a full random effects structure with random intercepts and slopes for languages and subfamilies. Due to the size of the data and complexity of the model, modeling was done in Julia rather than R. Random effects for language families were left out of the model because the Julia mixed models software does not allow more than two levels in a nested random effects structure. The level of family was dropped because it accounted for less variation than either subfamily and language.

Visualization of the data and residuals of the baseline model clearly indicated that curvature and/or interactions would be needed in the model. Therefore, I used an additive model selection process based on AIC to determine the final model. This process started with a baseline model with all three fixed effects but without interactions or curvature. At each step, the current model was compared to each of several alternative models

Formula	**AIC**	**Change from previous model**
Frequency ∼1 + SynND + OrthND + LevND + RE	239741	
Frequency ∼1 + poly(SynND, 2) + OrthND + SemND + RE	238103	Raise SynND to 2nd degree
Frequency ∼1 + poly(SynND, 3) + OrthND + SemND + RE	236332	Raise SynND to 3rd degree
Frequency ∼1 + poly(SynND, 3) + OrthND + SemND + poly(SynND, 3):SemND + RE	235528	Add interaction between SynND and SemND
Frequency ∼1 + poly(SynND, 4) + OrthND + SemND + poly(SynND, 4):SemND + RE	235214	Raise SynND to the 4th degree
Frequency ∼1 + poly(SynND, 4) + poly(OrthND, 2) + SemND + poly(SynND, 4):SemND + RE	234992	Raise OrthND to 2nd degree
Frequency ∼1 + poly(SynND, 4) + poly(OrthND, 3) + SemND + poly(SynND, 4):SemND + RE	234968	Raise OrthND to 3rd degree
Frequency ∼1 + poly(SynND, 4) + poly(OrthND, 3) + SemND + poly(SynND, 4):SemND + poly(OrthND, 3):SemND + RE	234859	Add interaction between SemND and OrthND

Table A.1: Model selection process from baseline model to final model, adding one polynomial degree or interaction at a time as licensed by AIC. Note that the random effects structure for each model, indicated by "RE", includes random intercepts and slopes for each fixed effect in that model (including polynomials and interactions).

	Est.	SE	z	p	language	subfamily
(Intercept)	-0.0755	0.0112	-6.74	<1e-10	0.0442	0.0354
syn_n_10_bc	-0.296	0.0273	-10.83	<1e-26	0.135	0.0791
syn_n_10_bc ^2	0.0829	0.0114	7.25	<1e-12	0.0608	0.0264
syn_n_10_bc ^3	0.016	0.0026	6.23	<1e-09	0.0087	0.0084
syn_n_10_bc ^4	-0.0051	0.001	-4.86	<1e-05	0.0047	0.0029
ft_n_10_bc	-0.3514	0.0221	-15.9	<1e-56	0.1064	0.0658
lev_n_10_bc	-0.1991	0.0197	-10.1	<1e-23	0.0519	0.0793
lev_n_10_bc ^2	0.0069	0.0036	1.92	0.0544	0.0084	0.0105
lev_n_10_bc ^3	0.0058	0.0026	2.27	0.023	0.0058	0.0095
syn_n_10_bc & ft_n_10_bc	0.0825	0.0141	5.87	<1e-08	0.035	0.0544
syn_n_10_bc ^2 & ft_n_10_bc	0.0018	0.0071	0.25	0.8041	0.0252	0.0232
syn_n_10_bc ^3 & ft_n_10_bc	-0.0053	0.0013	-3.9	<1e-04	0.003	0.0048
syn_n_10_bc ^4 & ft_n_10_bc	0.0009	0.0005	1.68	0.0929	0.0015	0.0018
lev_n_10_bc & ft_n_10_bc	0.0366	0.0104	3.51	0.0004	0.0385	0.0327
lev_n_10_bc ^2 & ft_n_10_bc	0.0094	0.0042	2.22	0.0262	0.0135	0.013
lev_n_10_bc ^3 & ft_n_10_bc	-0.004	0.0015	-2.6	0.0094	0.0038	0.0045
Residual	0.8239					

Table A.2: Coefficients for the final model predicting frequency from syntactic, semantic, and orthographic neighborhood densities.

that differed by the addition of either an interaction between two of the fixed effects or one degree of polynomial curvature for one of the fixed effects. (Any change in the fixed effects structure was mirrored in the random effects structure.) The best of the alternative models was chosen based on which one resulted in the greatest improvement to AIC, and the process was repeated until none of the alternative models offered an improvement on the current model. The models at each step of this process are shown in Table A.1. Table A.2 shows the model coefficients for the final model, as well as the residual variance.

A.2 Chapter 4 regression details

I fit a mixed-effects linear regression model predicting syntactic distance from semantic and orthographic distances. Both dependent and independent variables were Box-Cox

	Est.	**95% CI (lower)**	**95% CI (upper)**	**SD (family)**	**SD (subfamily)**	**SD (language)**
(Intercept)	0.001	-0.002	0.004	0.000	0.000	0.001
Orthographic distance	-0.007	-0.077	0.064	0.099	0.051	0.051
Semantic distance	0.301	0.260	0.342	0.045	0.047	0.057

Table A.3: Coefficients of the mixed-effects linear regression model predicting syntactic distance from orthographic and semantics distances for pairs of nouns in 48 languages

normalized and scaled within each language. The model included random intercepts and slopes (for both fixed effects) for the nested factors of language, subfamily, and family. For languages with more than 10,000 pairs, a sample of 10,000 rows was used in fitting the regression model. Data visualization and inspection of residuals did not suggest curvature for the model. Table A.3 shows the model coefficients, and the marginal R squared for this model is 0.088, while the conditional R squared is 0.110. I compared this model to two others, each with one of the fixed effects removed (although maintaining the same random effects structure). Unsurprisingly, the inclusion of semantic distance is justified ($\chi^2(1) = 30.1, p < .0001^{***}$), while the inclusion of orthographic distance does not significantly improve the model ($\chi^2(1) = 0.035, p = .852$).

A.3 Chapter 5 regression details

A mixed-effects logistic regression model was fit predicting gender sameness (*no* vs. *yes*) from syntactic, semantic, and orthographic distances and number of genders (a factor indicating whether the language has 2 or 3 genders). All three distance variables were Box-Cox normalized and scaled within each language. I included random intercepts for each language and language family, as well as random slopes for each of the fixed effects for both of these grouping levels. The regression was fit on randomly sampled parts of the data consisting of 10,000 pairs for each language. Data visualization and

	Est.	**95% CI (lower)**	**95% CI (upper)**	**SD (family)**	**SD (language)**
(Intercept)	0.202	0.060	0.344	0.000	0.002
Orthographic distance	-0.217	-0.315	-0.119	0.159	0.067
Semantic distance	-0.277	-0.387	-0.166	0.169	0.121
Syntactic distance	0.081	0.053	0.109	0.000	0.078
Number of genders (2)	-	-	-	0.153	0.148
Number of genders (3)	-0.745	-0.897	-0.592	0.012	0.096

Table A.4: Coefficients of the mixed-effects generalized linear regression model predicting gender sameness for pairs of nouns in 32 languages

inspection of residuals did not suggest curvature for the model. Table A.4 shows the model coefficients; the marginal R squared for this model is 0.056, while the conditional R squared is 0.080. I compared this model to four additional ones—each with one of the fixed effects removed (but no change to random effects)—using likelihood ratio test. These comparisons indicated that the full model explains the data better than ones without a fixed effect for semantic distance ($\chi^2(1) = 12.8, p < .001^{***}$), orthographic distance ($\chi^2(1) = 11, p < .001^{***}$), syntactic distance ($\chi^2(1) = 17.5, p < .0001^{***}$), and number of genders ($\chi^2(1) = 19.3, p < .0001^{***}$). I also fit a model with an interaction between syntactic distance and number of genders as an additional fixed effect, along with corresponding random slopes for language and family. A likelihood ratio test comparing this new model to one without the interaction (but no change to random effects) indicates that this interaction does not significantly improve the model ($\chi^2(1) = 0.025, p < .874$).

Bibliography

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order of acquisition affects what gets learned. *Cognition*, *122*(3), 292–305.

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

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

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313.

Baayen, R. H., Milin, P., Filipović-urević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482.

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in italian. *Perception Psychophysics*, *58*(7), 992–1004.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–76). Cambridge University Press.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Bickel, B. (2008). A refined sampling procedure for genealogical control. *STUF - Language Typology and Universals*, *61*, 221–233. https://doi.org/https://doi.org/10.1524/stuf.2008.0022

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Borer, H. (2005). *Structuring sense volume 1: In name only*. Oxford University Press.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience*, *35*(25), 9329–9335.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "bouba" and "kiki" in namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, *126*(2), 165–172.

Bresnan, J. (2001). *Lexical-functional syntax*. Blackwell.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin Review*, *8*, 531–544. https://doi.org/https://doi.org/10.3758/BF03196189

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge University Press.

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*(3), 348–369.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131–1157.

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

Chomsky, N. (1995). *The minimalist program*. MIT Press.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–508.

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.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th west coast conference on formal linguistics* (pp. 90–98). Cascadilla Proceedings Project.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Danguecan, A. N., & Buchanan, L. (2016). Semantic neighborhood effects for abstract versus concrete words. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01034

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

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2016). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149–2169.

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Dell, G. S., & Gordon, J. K. (2011). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 9–38). De Gruyter Mouton.

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

de Marneffe, M.-C., Manning, C., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*(2), 255–308.

de Saussure, F. (1916). *Course in general linguistics*. McGraw-Hill.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diessel, H. (2015). Usage-based construction grammar. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 295–321). De Gruyter.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2005). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language*, *13*, 257–292.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods*, *41*, 1210–1223. https://doi.org/https://doi.org/10.3758/BRM.41.4.1210

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Perspectives on morphological structure: Data and analyses* (pp. 212–239). Brill.

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.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2021). *Ethnologue: Languages of the world* (24th ed.). SIL International.

Edmiston, P., & Lupyan, G. (2015). What makes words special? words as unmotivated cues. *Cognition*, *143*, 93–100.

Enfield, N. J. (2014). *Natural causes of language: Frames, biases and cultural transmission*. Language Science Press.

Enfield, N. J. (2015). *The utility of meaning: What words mean and why*. Oxford University Press.

Faaß, G., & Eckart, K. (2013). Sdewac - a corpus of parsable sentences from the web. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.

Fillmore, C. (2015). The grammar of hitting and breaking. In R. A. Jacobs & P. A. Rosenbaum (Eds.), *Readings in english transformational grammar* (pp. 120–133). Ginn.

Fitneva, S., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, *36*(5), 967–997.

Flemming, E. (2004). *Contrast and perceptual distinctiveness. phonetically based phonology*. Cambridge University Press.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Futrell, R. (2010). *German grammatical gender as a nominal protection device* (Undergraduate Thesis). Stanford University. Stanford.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Gasser, M. (2005). The origins of arbitrariness in language. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 434–439). Cognitive Science Society.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

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

Graff, P. (2012). *Communicative efficiency in the lexicon* (Doctoral dissertation). Massachusetts Institute of Technology. Cambridge, MA.

Grainger, J., Muneaux, M., Farioli, F., & Ziegler, J. C. (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recogni-

tion. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *58A*(6), 981–998. https://doi.org/https://doi.org/10.1080/02724980443000386

Grosjean, F., Dommergues, J. Y., Cornu, E., Guillelmon, D., & Besson, C. (1994). The gender-marking effect in spoken word recognition. *Perception Psychophysics*, *56*(5), 590–598.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). *Glottolog 4.5*. Max Planck Institute for Evolutionary Anthropology.

Hausser, J., & Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*, 1469–1484.

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

Heath, J. (1975). Some functional relationships in grammar. *Language*, *51*(1), 128–149.

Hendrix, P., Bolger, P., & Baayen, R. H. (2017). Distinct erp signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 128–149.

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (1994). *Sound symbolism*. Cambridge University Press.

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

Hudson, R. (2007). *Language networks: The new word grammar*. Oxford University Press.

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

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65.

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

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), 630–645.

Kay, P. (2013). The limits of (construction) grammar. In T. Hoffmann & G. Trousdale (Eds.), *The oxford handbook of construction grammar* (pp. 32–48). Oxford University Press.

Kelly, M. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364.

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

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

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

Kostić, A., Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 1–44). Mouton de Gruyter.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Langacker, R. W. (1987). *Foundations of cognitive grammar, vol. i: Theoretical prerequisites*. Stanford University Press.

Lester, N. (2018). *The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition* (Doctoral dissertation). University of California Santa Barbara. Santa Barbara, CA.

Lester, N., Auderset, S., & Rogers, P. (2018). Case inflection and the functional indeterminacy of nouns: A cross-linguistic analysis. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2029–2034.

Lester, N., Feldman, L. B., & Moscoso del Prado Martín, F. (2017). You can take a noun out of syntax...: Syntactic similarity effects in lexical priming. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2537–2542.

Lester, N., & Moscoso del Prado Martín, F. (2015). Constructional paradigms affect visual lexical decision latencies in english. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 1320–1325.

Lester, N., & Moscoso del Prado Martín, F. (2016). Syntactic flexibility in the noun: Evidence from picture naming. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2585–2590.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (pp. 849–856). MIT Press.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. https://aclanthology.org/D08-1025

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala & J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Academic Press.

Linzen, T., Marantz, A., & Pylkkänen, L. (2013). Syntactic context effects in visual word recognition. *The Mental Lexicon*, *8*, 117–139.

Locker, L., Simpson, G., & Yates, M. (2003). Semantic neighborhood effects on the recognition of ambiguous words. *Memory Cognition*, *31*, 505–515. https://doi.org/ https://doi.org/10.3758/BF03196092

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*, 1246.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, Computers*, *28*(2), 203–208.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186.

Macdonald, G. (2013). *Aging and semantic processing* (Doctoral dissertation). University of Windsor. Windsor, ON.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Wordforms are structured for efficient use. *Cognitive Science*, *42*(8), 3116–3134.

Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics*, *4*, 201–225.

McDonald, S., & Shillcock, R. (2001a). Contextual distinctiveness: A new lexical property computed from large corpora.

McDonald, S., & Shillcock, R. (2001b). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295–322.

McRae, K., Cree, G. S., Seidenberg, M. S., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559. https://doi.org/https://doi.org/10.3758/BF03192726

Mel'čuk, I. (1988). *Dependency syntax: Theory and practice*. The SUNY Press.

Miestamo, M., Bakker, D., & Arppe, A. (2016). Sampling for variety. *Linguistic Typology*, *20*(2), 233–296. https://doi.org/https://doi.org/10.1515/lingty-2016-0006

Milin, P., Filipović-urević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, *60*, 50–64.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Cognition*, *34*(1), 65–79. https://doi.org/10.1037/0278-7393.34.1.65

Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 139–164). John Benjamins.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*, *369*(1651), 20130299.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182.

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

Monaghan, P., Maddock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1152–1164.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*, 1–18.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of speech, language, and hearing research*, *47*(5), 1048–1058. https://doi.org/https://doi.org/10.1044/1092-4388(2004/078)

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, Computers*, *36*, 402–407. https://doi.org/https://doi.org/10.3758/BF03195588

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit representations. *Psychological Review*, *99*(2), 322–348. https://doi.org/https://doi.org/10.1037/0033-295X.99.2.322

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

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, *4*(2), 115–125.

Nivre, J. (2005). *Dependency grammar and dependency parsing* (Technical Report MSI report No. 05133). Växjö University: School of Mathematics and Systems Engineering.

North, B. V., Curtis, D., & Sham, P. C. (2002). A note on the calculation of empirical p values from monte carlo procedures. *American Journal of Human Genetics*, *71*(2), 439–441.

Nowak, M. A., Krakauer, D. C., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society B: Biological Sciences*, *266*(1433), 2131–2136.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*(1), 181–186.

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.

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

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. The University of Chicago Press.

Rafferty, A. N., & Manning, C. D. (2008). Parsing three german treebanks: Lexicalized and unlexicalized baselines. *In ACL Workshop on Parsing German*.

Ramchand, G. C. (2008). *Verb meaning and the lexicon: A first phase syntax*. Cambridge University Press.

Raymond, W. D., Dautricourt, R., & Hume, E. (2006). Word-internal /t, d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change*, *18*(1), 55–97.

Reilly, J., Westbury, C., Kean, J., & Peelle, J. E. (2012). Arbitrary symbolism in natural language revisited: When word forms carry meaning. *PloS one*, *7*(8), e42286.

Rijkhoff, J., & Bakker, D. (1998). Language sampling. *Linguistic Typology*, *2*, 263–314. https://doi.org/https://doi.org/10.1515/lity.1998.2.3.263

Robbeets, M. I. (2017). Austronesian influence and transeurasian ancestry in japanese: A case of farming/language dispersal. *Language Dynamics and Change*, *7*, 201–251.

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

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 841–850.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavioral Research Methods*, *42*, 393–413. https://doi.org/10.3758/BRM.42.2.393

Shillcock, R., Kirby, S., McDonald, S., & Brew, C. (2001). Filled pauses and their status in the mental lexicon. *Disfluency in Spontaneous Speech (DiSS'01)*, 53–56.

Siakaluk, P. D., Buchanan, L., & Westbury, C. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory Cognition*, *31*(1), 100–113. https://doi.org/https://doi.org/10.3758/BF03196086

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, *39*(1), 212–226.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386.

Stabler, E. P. (2013). Two models of minimalist, incremental syntactic analysis. *Topics in Cognitive Science*, *5*, 611–633.

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

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

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211.

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32*(4), 827–853.

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*, *49*(6), 1175–1192.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132.

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*(2), 259–278.

Tamariz, M., & Kirby, S. (2015). Culture: Copying, compression, and conventionality. *Cognitive Science*, *39*(1), 171–183.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467.

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

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of italian tongues. *Psychological Science*, *8*(4), 314–317.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. https://doi.org/https://doi.org/10.1016/j.cogpsych.2003.09.001

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747.

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

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

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

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.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*(2), 179–186.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*, 1272–1288.

Williams, A., Blasi, D., Wolf-Sonkin, L., Wallach, H., & Cotterell, R. (2019). Quantifying the semantic core of gender systems. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5734–5739.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449.

Yates, M., Locker, L., & Simpson, G. B. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory Cognition*, *31*(6), 856–866. https://doi.org/https://doi.org/10.3758/BF03196440

Yates, M., Locker Jr., L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin Review*, *11*(3), 452–457. https://doi.org/10.3758/bf03196594

Zipf, G. K. (1935). *The psycho-biology of language*. Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Zubin, D., & Köpcke, K.-M. (1986). Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In C. Craig (Ed.), *Noun classification and categorization: Proceedings of a symposium on categorization and noun classification, eugene, oregon, october 1983* (pp. 139–180). John Benjamins.