Data-driven Crosslinguistic Modeling of Constituent Ordering Preferences

By

Zoey (Ying) Liu
Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Kenji Sagae, Chair

John A. Hawkins

Emily Morgan

Committee in Charge

2020

To my grandpa

Contents

List of Figures

List of Tables

Abstract

Data-driven Crosslinguistic Modeling of Constituent Ordering Preferences

Why are languages the way they are? In this dissertation, I take up this question with a focus on crosslinguistic constituent orderings. Specifically, borrowing insights from language processing and language evolution, I ask what abstract constraints as well as idiosyncratic biases govern language users' choice among grammatical alternatives of the same syntactic constructions across genres and languages. Adopting a data-driven approach, I explore three directions in particular. First, from Chapter 3 to Chapter 6, taking advantage of large-scale multilingual corpora, I investigate and quantify the roles of numerous factors that are motivated by long-standing linguistic theories as well as previous empirical findings in word order preferences. I show that while the effect of individual factors depends on the ordering structures of different languages, generally the predictive power and direction of these constraints are more dependent on whether the orderings are in the preverbal or the postverbal domains. In addition, besides these abstract constraints that yield probabilistic typological tendencies, in Chapter 7 I ask why language users have idiosyncratic ordering preferences and how regularization of this idiosyncrasy arises diachronically, using Bayesian iterated learning models that simulate the process of language change. Lastly, I adopt the theoretical framework of dependency syntax to develop a dependency treebank for Hupa, an endangered Dene language of northwestern California, as a way to formalize and model the syntax of indigenous languages. [1]

[1] Codes and data for experiments in this dissertation are available at `https://github.com/zoeyliu18/Dissertation`.

Acknowledgments

Unbounded gratitude, parsed in linear time, to my advisor Kenji Sagae. Thank you, Kenji, for telling me about “The Karate Kid", for encouraging me in your own way, for leading me to develop my own research style, for inspiring me that one does not have to do what is considered *popular*, for supporting me in every aspect of education, and for being my advisor. I am passionate about what I do; I believe in what we do as a community of educators and scholars. And that is because of you.

I would like to thank my other two wonderful committee members, Jack Hawkins and Emily Morgan. Jack, my dissertation has benefited tremendously from your insights, your knowledge, and your continuous encouragement. Emily, I am grateful for your feedback in both writing papers and giving conference presentations. I would not have learned about approaches from the language evolution literature (as soon as I did) if it were not for you.

Many other professors have had an impact on the way I think about research along the journey. Thank you to: Raúl Aranovich, for telling me that one could always do more analyses; Santiago Barreda, for showing me p value means nothing but one could still be statistically significant; Vai Ramanathan, for knowing how to make a great chicken pot pie; Justin Spence, for introducing me to the humanitarian aspects in research; Fernanda Ferreira, for always being kind to me.

Graduate school would not have been as fun without my beautiful friends. Kayla, I miss being goofy; Cesar, I miss procrastinating; Tyler, I miss seeing you just across the hall; Cari, I miss not being able to see you just across the hall; Sam, I miss our “fighting" over GPU space; Tom, I miss your pumpkin pie; Jeff, I miss slow cooker; Xinye, I miss you saving me food; Cherise, I miss eating hotpot; Lauren, I miss getting boba; Skyler, I miss taco Tuesday; Zion, I miss our Thanksgiving; Brennan, I miss our Trader Joe’s trips. Music, my best friend, letting me have my own Octopus’s Garden. To my persons: Alyx, my close confidant; Annie, my travel companion; Crystal, my strengths of happiness.

My gorgeous mama and my dapper dad, I love you. You give me purpose. Thank you for always believing I could do anything, even though you know I am a terrible dancer. I will tell you the rest in Mandarin when I see you.

Chapter 1

Introduction

Languages differ appreciably yet share considerable similarities across a wide range of structures. Characterizations and quantifications of these crosslinguistic regularities as well as differences can not only further our fundamental understanding of languages, but also provide findings with practical implications for the development of language technologies. The availability of multilingual corpora annotated with syntactic information allows for investigation of issues in linguistic typology using computational methods (Merlo, 2015; Wang and Eisner, 2017, 2018a,b; Cotterell and Eisner, 2018; Östling, 2015; Cotterell et al., 2019), shedding new light on the underlying mechanisms behind structural patterns in natural languages. In addition, typological information can help improve the performance of multilingual natural language processing systems (Bender, 2009; Ponti et al., 2019) and facilitate the extension of current computational techniques and applications from resource-rich languages to low-resource or understudied languages (O'Horan et al., 2016).

Certain structural patterns have been found to occur both within and across languages. For instance, people tend to use warm color terms such as red and yellow to talk about the main topic or object of conversation, while they resort to cold color terms like green and blue when describing background information (Gibson et al., 2017). In most attested languages, subjects tend to appear before objects (Dryer, 2013). Languages with oblique phrases occurring mostly before the head verb are likely to have head-final relative clauses (Lawyer, 2015).

On the other hand, linguistic variation occurs at diverse levels of granularity (e.g. lexical, morphological, phonological, syntactic, etc). Besides these general patterns, structural idiosyncrasies also exist both within and across languages. For example, although the order of subject (S), object (O) and verb (V) in English predominantly follows SVO, having evolved from the dominant SOV order of Old English (Tily, 2010), English also contains instances following an OSV order, such as *The book Cat's Cradle, I like*, possibly due to different social or pragmatic motivations (Prince, 1988). While the ordering of items in the noun phrases (NP) in Italian largely follows Det > Num > Adj > N, the most common NP order in Thai shows the mirror image (N > Adj > Num > Det) (Martin et al., 2019). Despite the fact that the general classifier 个 in Mandarin Chinese can be used with a wide range of nouns belonging to different semantic classes, certain nouns tend to be paired with specific classifiers (e.g. 熊猫 (*panda*) with the classifier 只 (*zhi*)) (Liu et al., 2019).

These general structural characteristics as well as idiosyncratic observations result in a set of typological patterns that are probabilistic and gradient in nature (Ponti et al., 2019), rather than deterministic or absolute (Greenberg, 1963). So what gives rise to these typological patterns in the first place? Previous research has tried to address this broad question from a variety of angles. From the perspective of generative grammar (Chomsky, 1968), language users all have innate or biological constraints; they could simply tune their "parameter settings" (Reali and Christiansen, 2009) to adjust to the surrounding environment, which leads to both similarities and differences in their language usage. From the perspective of sociolinguistics, socially and cognitively motivated interactions as well as dynamics between speakers from different speech communities shape language structure and ultimately language change (Labov, 2011). From the perspective of processing, language variation is motivated by the pressure to achieve communication efficiency and processing ease, which in turn defines optimal linguistic forms and functions (Hawkins, 2014; Hahn et al., 2020; Blasi et al., 2019). From the perspective of language evolution, structural variation arises from the continuous process of cultural transmission and language learning (Christiansen and Chater, 2008; Christiansen and Kirby, 2003).

1.1 Syntactic variation as a multifactorial function

Among the different types of structural tendencies, variation in syntactic patterns has perhaps attracted the most amount of attention. In particular, there has been a great deal of research devoted to providing empirical accounts for why word orders are the way they are (Futrell et al., 2015a; Hawkins, 1994; Culbertson et al., 2012; MacDonald, 2013; Gibson and Wu, 2013; Wasow and Arnold, 2003; Jaeger and Tily, 2011; Jaeger and Norcliffe, 2009; Christianson and Ferreira, 2005; Christiansen and Chater, 2008).

Within the domain of syntactic ordering preferences, cases are ubiquitous where language users can express the same semantically invariant idea using different grammatical alternatives of the same syntactic construction. As an illustration, consider the following examples from Hupa (Sapir and Golla, 2001), an endangered Dene language of northwestern California, traditionally spoken in Hoopa Valley on the lower Trinity River in present-day Humboldt County. Though Hupa typically has an SOV order, there are also relatively frequent instances where the subject or the object occurs freely after the head verb. When the subject of the sentence *hay tsumehstł'o:n (the woman)* moves from before the verb, as in (1), to a postverbal position, as in (2), the sentence still has the same meaning and both alternatives are grammatical.

(1) hay tsumehstł'o:n na'te:dichwiw
the woman she cried along

(2) na'te:dichwiw hay tsumehstł'o:n
she cried along the woman
'The woman cried as she went back along.'

Now it is natural to ask: what linguistic principles govern the observed ordering choices? Previous research has pointed to various structural constraints that influence constituent orderings, including but not limited to: syntactic weight (Wasow, 1997b), semantic closeness/argument status (Culicover and Jackendoff, 2005), discourse status (Arnold et al., 2000), animacy (Bock and Warren, 1985), and verb subcategorization frames (Wasow and Arnold, 2003), among many others. Hawkins (2014) proposed that crosslinguistic syntactic variation is a multifactorial process, that word order preferences are motivated by several competing and cooperating factors simultaneously.

1.2 A preview of contributions in this dissertation

Taking a bird's eye view of the field, it is the case that most previous studies on syntactic alternations have focused on a limited set of factors in a limited set of languages (Temperley, 2007; Gildea and Temperley, 2007; Park and Levy, 2009; Wasow and Arnold, 2003; Yamashita and Chang, 2001). Their findings are not directly comparable as most experiments have examined different syntactic constructions. These limitations mean that it is not currently clear what the best typological determinants are for constituent orderings across languages. Whether certain factors play a comparable role crosslinguistically or whether they operate differently within and across languages remains unknown. Thus the proposal that crosslinguistic syntactic variation is a function of multiple factors still lacks proper empirical support.

Investigating structural constraints provides a lens into the knowledge that language users have about the constructions of their language and it enables us to probe how they comprehend and produce linguistic expressions. From a usage-based perspective (Bybee, 1985; Ellis, 2002; Cameron-Faulkner et al., 2003; Gries and Ellis, 2015; Gries, 2005; MacWhinney et al., 2014), this dissertation tests and explores numerous factors to ultimately explain regularities and differences in crosslinguistic constituent ordering preferences. Overall, this dissertation demonstrates that both the extent and direction of predictions by structural constraints depend on whether the ordering is in the preverbal or postverbal domains.

More specifically, taking a data-driven approach, I leverage multilingual corpora and computational measures. Borrowing insights from both language processing (Hawkins, 2004) and language evolution (Kirby et al., 2014). I characterize and quantify how different lexical, semantic and syntactic-level factors as well as item-specific idiosyncrasies predict constituent orders at a large scale. I show that multiple constraints are at play in order to derive the ideal surface ordering structure, providing direct evidence that syntactic variation is indeed a multifactorial function of competing and cooperating motivations. With the studies presented here, I hope to bring new data and insights to bear on existing theoretical principles, and to pave a methodological path forward for future research

directions in language typology.

In details, Chapter 2 describes the background information for the major structural factors and datasets to be explored in this dissertation. Chapter 3 tests the prediction of the Dependency Length Minimization (DLM) principle (Temperley and Gildea, 2018). Specifically, in this chapter, I raise two general questions regarding the effect of dependency length in ordering preferences. First, is there a typological tendency for shorter constituents to appear closer to their syntactic heads in constructions with flexible orderings? Secondly, how does the extent of DLM in these constructions vary for languages with different structural characteristics? I take up these questions using prepositional and postpositional phrase (hereafter both referred to as PP) ordering as a test case in 34 languages taken from the Universal Dependencies project version 2.5 (UD) (Zeman et al., 2019). I focus on sentences with verb phrases (VP) that have exactly two PP dependents on the same side of the head verb, the ordering of which permits flexibility in at least some contexts. Overall I show that there is a pronounced preference for shorter PPs to be closer to the head verb. This is the first large-scale quantitative evidence showing that DLM is also found in syntactic alternations within many different languages (as opposed to across languages, cf. e.g., Futrell et al., 2015).

Chapter 3 also shows that while the strength and applicability of DLM depend on the specific orderings and structures of particular language types. in general there appears to be a much stronger preference for DLM across languages when the two PPs appear after the head verb, compared to no or a much weaker tendency for shorter dependencies when the two PPs are before the verb. This contrast is most visible in mixed-type languages with head-initial PPs that can appear both after or before the head verb. In the limited number of rigid OV languages in the dataset, which have head-final PPs before the head verb, I observe no robust tendency for DLM, in contrast to the patterns in languages with head-initial PPs after the head verb. This is at variance with previous findings of a long-before-short preference in preverbal orders of head-final languages (Japanese (Yamashita and Chang, 2001; Yamashita, 2002); Korean (Choi, 2007)).

Findings from Chapter 3 call for investigations of other structural constraints that

operate together with dependency length, which possibly results in the contrasting preference for DLM between preverbal and postverbal domains. Chapter 4 makes an attempt to address this with a focused comparative analysis of constituent orders in English and Mandarin Chinese, two languages with distinct typological word order characteristics. Using the Penn Treebank (Marcus et al., 1993) for English and the Penn Chinese Treebank version 5 (Xue et al., 2005) for Chinese, I examine the effects and the interactions of three factors: dependency length, argumenthood status and the adverbial ordering rule, Manner Place Time, using PP ordering again as the test case. These two languages have different PP ordering structures. When a head verb has two PP dependents to the same side, they appear as head-initial PPs after the verb in English; whereas they occur as head-initial PPs before the verb in Chinese. I show that while dependency length plays a strong role in postverbal domains across genres for English, it exerts only a mild effect in preverbal orders in Chinese. On the other hand, the argument status of the PP has a pronounced role in both languages. There exists a strong tendency for the argument-like PP to appear closer to the head verb than the adjunct-like PP in both preverbal and postverbal orderings.

The effects for both argument status and Manner Place Time measured in Chapter 4 rely on the gold standard manual annotations in the data, which are not available directly for multilingual corpora at a much larger scale. To explore the roles of different structural factors including argument status in a crosslinguistic context, in Chapter 5, I borrow techniques from computational linguistics and natural language processing. I show how to computationally measure the effects of semantic closeness, lexical frequency, contextual predictability and word co-occurrence information. The latter three have been found or suggested to be good predictive constraints in psycholinguistic processing, though their various roles have not been examined much in syntactic alternations. Specifically, I compute the effects of the aforementioned four factors using word embeddings, estimates from unigram language models, estimates from neural contextual language models, and pointwise mutual information, respectively.

Another notion discussed in Chapter 3 is the relationship between ordering flexibil-

ity and the extent of DLM. In line with a previous observation by Futrell et al. (2015a) and Gildea and Temperley (2010), languages with longer dependencies *seem* to also be those that have been claimed to contain more word order freedom. Whether that statement is quantitatively valid or not remains unclear, however. Chapter 6 provides a simple way to explore the relationship between word order flexibility and dependency length. The test case is syntactic construction in which the head verb has a direct object NP dependent and exactly one PP dependent adjacent to each other appearing on the same side. I use data from UD and approximate ordering flexibility with *entropy* (Futrell et al., 2015b; Levshina, 2019), a measure of dispersion. Results for a total of 36 languages reveal an overall negative correlation between degree of word order freedom and the effect for dependency length in predicting ordering preferences. This indicates that in domains with more ordering variability, there is a weaker tendency for shorter dependencies.

The experiments of Chapter 3 to Chapter 6 have focused on the explanatory power of various abstract constraints in distinct ordering patterns. Particularly they have tried to compare and account for the different ordering preferences in preverbal and postverbal structures. But how important is the specific verb in predicting syntactic choices? Previous studies have noted the significant effects of item-specific knowledge and idiosyncrasies (Goldberg, 2003, 2009; Morgan and Levy, 2015, 2016a,b; Ferreira, 1994) in constituent orders. These studies have demonstrated that language users can have idiosyncratic preferences when it comes to word orders. However, the role of idiosyncrasy has not been subjected to serious investigations in syntactic alternations.

In Chapter 7, I examine how item-specific knowledge of individual verbs in syntactic constructions affects constituent ordering preferences. Using the dative construction in English as a test case, I show that the more frequent a construction type is, distinguished by the head verb, the stronger and more extreme the preference is that language users will have for one alternative over the other, a pattern that is known as *regularization* in statistical learning. Adopting methodology of iterated learning models from the literature of language evolution and change (Reali and Griffiths, 2009; Kirby et al., 2014), I demonstrate further that regularization arises from the interactions between language production

as well as the continuous process of cultural transmission and language learning.

Finally, most of the languages studied in this dissertation are Indo-European. That also seems to be the case throughout the language sciences. Chapter 8 reports on an effort to adapt theoretical framework of dependency grammar to develop a dependency syntactic treebank for Hupa (Sapir and Golla, 2001) as a way to formalize and model the syntax of indigenous languages. I discuss the potential benefits of developing such corpus annotations for language documentation in the hope that the field will actually start to become more inclusive and pay significant attention to endangered and low-resource languages.

Chapter 2

Background

2.1 Datasets and annotation scheme

This dissertation takes advantage of three datasets: the Penn Treebank (PTB) (Marcus et al., 1993), the Penn Chinese Treebank version 5 (CTB) (Xue et al., 2005), and multilingual corpora from the Universal Dependencies project version 2.5 (UD) (Zeman et al., 2019). Both PTB and CTB are annotated using a phrase-structure format. PTB includes three corpora annotated with syntactic constituency structures that cover both written and spoken domains in English. Specifically, it contains approximately one million words of text in English from each of: the Wall Street Journal (WSJ), the Brown corpus (Kučera and Francis, 1967) and transcriptions of spontaneous spoken conversations from the Switchboard corpus (Godfrey et al., 1992). On the other hand, CTB has a total of 500K words.

By contrast, UD uses the dependency grammar theoretical framework. The historical origins of dependency grammars can be traced back to Tesnière (1959). Under this framework, phrasal structures are not encoded directly. Instead, dependency structure is heavily lexicalist, in that it focuses on individual words, or morphological lemmas.

In the annotation scheme of UD, within a sentence, or a clausal structure, every word is in a head-dependent pair. The head and the dependent are connected by a binary dependency relation. Here the dependency relation between each head and its dependent is indicated on the arrow, pointing from the head to its dependent. One

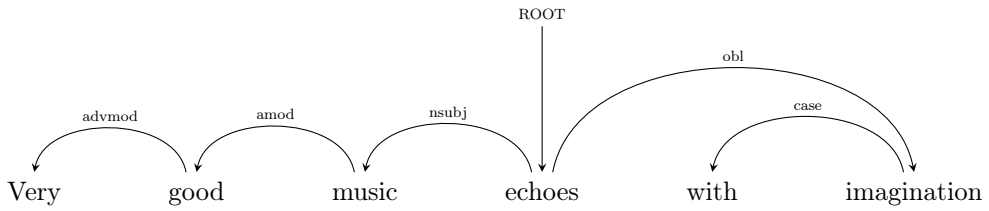


Figure 2.1: An example of dependency structure representation based on the UD annotation scheme.

dependent can only have one head while one head can have multiple dependents.[1] The dependency representations tend to be verb-centric, with the lexical head verb being the center of a clause or the *root* of a sentence in most cases. For instance, as shown in Figure 2.1, the verb *echoes* is the *root* of the sentence. The verb has two dependents, *music* and *imagination*, and their dependency relation with the verb is *nsubj* (subject) and *obl* (oblique), respectively.

The representation in Figure 2.1 can be easily transformed to that in Figure 2.2, a hierarchical dependency tree (Gulordava, 2018) which offers a straightforward view as to which word is the root of the sentence, what words are the children and the grandchildren of the word *echos*, for example. This allows dependency structure to step away from the traditional leaf nodes in the phrasal structure, and creates constituent structure by combining head-dependent pairs to construct dependency subtrees. For instance, instead of having an NP to represent the constituent *Very good music*, the phrase is built by connecting together the head noun *music* and *good*, as well as *good* and *very*, via their corresponding dependency relations.

The representations of dependency structure make it possible to better capture the syntax of languages with relatively free word orders (Jurafsky and Martin, 2019) and discontinuous constituency structures. As an illustration, let us take a look at the examples in German in Figure 2.3. The two sentences have the same semantic meaning, expressed in different syntactic orderings. The grammatical functions of the NPs are denoted by

[1]This is the case in the "basic" representations of UD (de Marneffe and Nivre, 2019). In the "enhanced" representations (Schuster and Manning, 2016), words in certain syntactic constructions can have more than one head. The studies in this dissertation consider only the "basic" representations, which are more widely applied across the languages in UD.

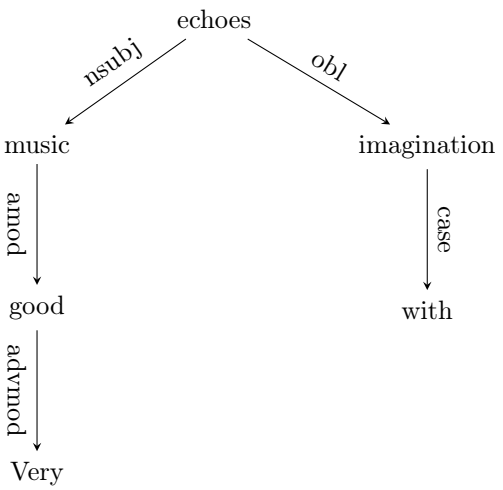


Figure 2.2: A hierarchical dependency tree representation transformed from Figure 2.1.

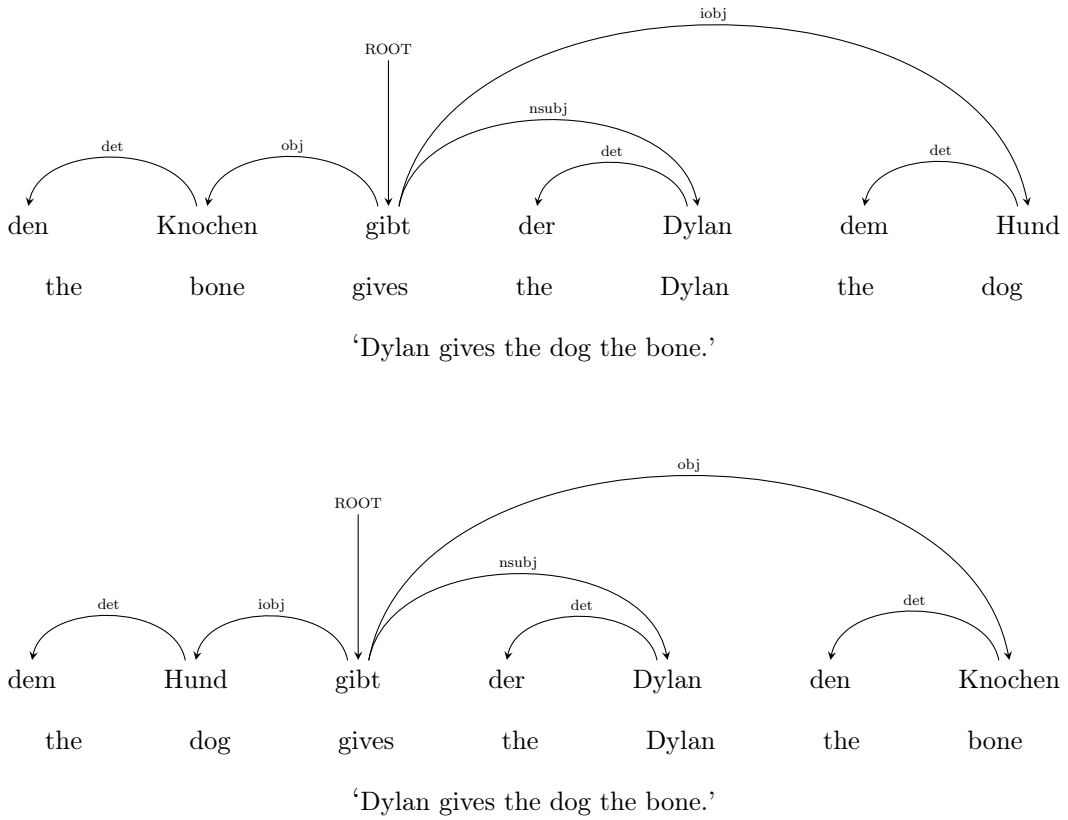


Figure 2.3: Demonstrations that dependency representations are able to better capture the syntax of languages (e.g. German) with more word order flexibility.

case-marking the determiners, yielding flexible permutations of the subject, direct and indirect object within the sentence. Here neither of the two sentences can be represented by a right-branching or left-branching phrase-structure tree. However, the dependency representations can depict these grammatical functions while still preserving the original linear word orders.

Overall, the goal of UD is to provide consistent and gold standard annotations for typologically diverse languages. At the same time it also allows adaptations to language-specific properties. Currently UD uses 17 part of speech (POS) tags and 37 universal types of dependency relations in total, with more syntactic relation subtypes defined for specific languages. At the time of writing this dissertation, UD contains 157 treebanks for 90 languages. Each treebank is a syntactically parsed and annotated textual corpus, and certain languages have more than one treebank available. In the annotation format of each treebank, each token within each sentence has its own index. Information regarding the lemma, POS, syntactic head, dependency relation and sometimes other morphological features such as the gender and case of each token is explicitly encoded. Among the three datasets, UD is the main one used in this dissertation, and the analyses to be given here derive largely from dependency structures.

The annotation system in UD favors content words over function words as the syntactic heads of most of the dependency relations within the sentence. In this annotation system, for instance, within an adposition phrase, the lexical noun is treated as the head of the adposition (*Lady* is the head of *of* in *The Portrait of a Lady*); the lexical verb is considered the head of complementizers/subordinators and auxiliaries (*write* is the head of *that* and *should* in *It is true that I should write my dissertation*); the copula verb *is* in the sentence *The wondrous life of Oscar Wao is brief* is the dependent of *brief*. This specific annotation scheme distinguishes UD from other syntactic frameworks (e.g. Government and Binding theory (Chomsky, 1993); Head-driven phrase structure grammar (Pollard and Sag, 1994); the Minimalist Program (Chomsky, 1995); Lexical Functional grammar (Bresnan, 2001)) and from standard theoretical assumptions (Abney, 1987) in the literature, according to which, for example, adpositions are heads of PP, not dependents.

This annotation choice has drawn a thorough critique from Osborne and Gerdes (2019). Especially from a typological perspective, they argued that the current scheme of UD makes languages such as English and French more head-final than head-initial. To support their argument, they gave the following example in English (Figure 2.4), where + represents the dependent appearing before its syntactic head, and - the dependent appearing after. In this example, there are five dependents that occur before their heads (in bold) and two that occur after, making more head-final structures than head-initial ones.

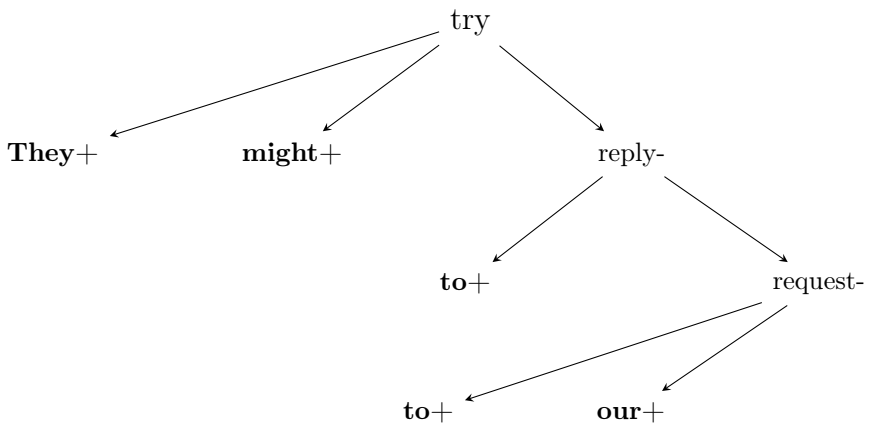


Figure 2.4: A demonstration of calculations for the number of head-initial and head-final structures in Osborne and Gerdes (2019). Dependents that occur before their syntactic heads are in bold.

Respectfully, I actually disagree with their argument here. The precise definition of headedness as well as the categorization of a language as head-initial, head-final, or rather, mixed-type are still not exactly clear (Polinsky, 2012). Though many have assumed English to be *consistently* head-initial, this is in practice not the case, considering the order of adjective and head nouns for example. The overall headedness of a language should depend on the precise definition of a syntactic head, on the headedness of syntactic constituency at different levels, and accordingly on the relative ratio of head-initial to head-final structures.

Following this line of thought, given the dependency representation above, one way to calculate headedness would be as follows. There are three dependency subtrees, or constituents that will be taken into consideration here: the subtrees headed by *try*, *reply*

and *request*, respectively. Within the subtree of *try*, there are two pre-dependents and five post-dependents, making this constituency head-initial overall. Within the subtree of *reply*, there is one pre-dependent and three post-dependents, again making this constituent head-initial. Within the subtree of *request*, there are two pre-dependents, making this constituent head-final. Overall two out of the three constituents are head-initial, meaning that the full sentence can be defined as head-initial as well.

In addition, giving priority to the content head within a syntactic construction is not to say that the content head is the *right* head of the construction or *should* be the head, as there is still much debate regarding the treatment of certain structures such as coordination (de Marneffe and Nivre, 2019). Rather, the motivation behind this choice is in order to provide comparable structural representations for different morphosyntactic constructions when they have similar and/or comparable grammatical functions or semantic meanings. For example, the locative PP *at the house* in English consists of the preposition *at* and the lexical noun *house*. The same meaning in Walpiri, a Pama-Nyungan language spoken in Australia, is expressed as an NP consisting of one noun *yuwarlirla*, in which a locative case marker *-rla* is attached to the word for house, *yuwarli* (Bavin, 1990). The noun in each of the two phrases (*house* in English, *yuwarlirla* in Warlpiri) can legitimately be regarded as the head in this example, for reasons of crosslinguistic consistency and comparability. Similarly, the temporal phrase *por la mañana* in Spanish can be expressed as either a nominal, 早上, or a PP 在早上 in Mandarin Chinese. These phrases all have the content word meaning *morning* here, but they don't always have the preposition.

Overall, the annotation system of UD may be far from perfect, but the corpora provided do offer valuable resources for the typological analysis of syntactic variation, and it does make possible an alternative definition of (degrees of) headedness, which is useful for crosslinguistic comparison even though it may be at variance with some current and widely held theoretical assumptions.

With the data described above, I address the multifactorial nature of crosslinguistic syntactic variation via testing the predictions of a wide range of structural constraints

motivated by both theoretical principles and indications from existing quantitative evidence. In the following section, I review and discuss related work on the major factors to be explored in this dissertation.

2.2 Structural constraints

2.2.1 Dependency length

The role of dependency length in constituent orderings has been addressed in different principles, including Early Immediate Constituents (EIC) (Hawkins, 1994), Minimize Domains (MiD) (Hawkins, 2004), Dependency Locality Theory (DLT) (Gibson, 1998) and Dependency Length Minimization (DLM). While EIC, MiD and DLT were all built upon a phrase structure framework, DLM was formulated using the dependency grammar framework. Though EIC, MiD and DLT have not used the term *dependency length* directly, the notion can be easily derived. These principles are all based on similar empirical support and motivation, mainly from language comprehension studies, to the effect that shorter dependencies are preferred in order to ease processing and efficient communication (Jaeger and Tily, 2011; Gibson, 2000; Hawkins, 2014). While DLM has gained wide popularity over recent years (White and Rajkumar, 2012; Rajkumar and White, 2014; Futrell et al., 2015a; Liu, 2010), it is crucial to be familiar with the other three related theoretical principles as well.

2.2.1.1 Early Immediate Constituents (EIC) and Minimize Domains (MiD)

First proposed in Hawkins (1990, 1994), EIC suggests that the earlier that comprehenders recognize the syntactic structure of the linguistic sequence that is being parsed, or the *Phrasal Combination Domain (PCD)*, the easier it is for them to process this sequence.

> The human processor prefers linear orders that minimize PCDs (by maximizing their IC-to-word ratios), in proportion to the minimization difference between competing orders. (cf. Hawkins 2014: 12)

Extending EIC, Hawkins (2004) proposed a more generalized and inclusive principle for syntactic ordering preferences, *Minimize Domains (MiD)*, which suggests that not only

syntactic domains but also the semantic and lexical domains all need to be minimized in order to facilitate processing for the human language parser:

> The human processor prefers to minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed. The degree of this preference is proportional to the number of relations whose domains can be minimized in competing sequences or structures, and to the extent of the minimization difference in each domain. (Hawkins 2004: 31)

Consider the following examples with respect to syntactic aspects of the two principles:

(1) He $[_{VP}$ wrote 1984 $[_{PP_1}$ in three years $]$ $[_{PP_2}$ on a sparsely populated island $]]$.

(2) He $[_{VP}$ wrote 1984 $[_{PP_1}$ on a sparsely populated island $]$ $[_{PP_2}$ in three years $]]$.

In both sentences, the VP has two PP constituents, shown within square brackets. Switching the order of the two PPs results in alternatives that are both grammatical and convey the same meaning. In order for comprehenders to parse the structure of the VP, its immediate constituents must be identified. In (1), the verb, PP_1 and PP_2 can be recognized on the basis of six words from *wrote* to *on*, compared to the eight words required in (2), where the longer PP is placed before the shorter one. In other words, putting the shorter PP in the final position of the sentence postpones the recognition of the second preposition, and consequently the projection to the mother node of the second preposition, PP_2, thereby delaying the construction of the whole phrase. Thus based on EIC and MiD, a short-before-long preference is expected for postverbal constituent orderings, whereas the opposite long-before-short ordering tendency is predicted and observed for preverbal word orders (cf. Hawkins 2014: 96-98 for a summary). The motivation for these two principles is rooted in online parsing efficiency. Compared to an alternative structure that expresses the same meaning, if a phrasal structure can be derived from a smaller number of words in parsing, this means that the processing demands required for this structure decrease, working memory load is less taxed, and there is greater parsing efficiency.

2.2.1.2 Dependency Locality Theory (DLT)

Gibson (2000) proposed DLT as a theory of resource use constrained by memory limitations in language comprehension. According to this principle, the syntactic complexity of a sentence is composed of two factors: "storage cost", the cost of maintaining in memory the structure that has been built so far, and "integration cost", the cost of integrating, or matching the current word to previous words and structures that the current one contracts syntactic relations with.

> Resources are required for *two* aspects of language comprehension: (1) *storage* of the structure built thus far [...] and (2) integration of the current word into the structure built thus far. (p. 102)

Integration cost is where dependency length, or distance comes into play:

> The structural integration complexity depends on the *distance* or *locality* between the two elements being integrated. (p. 102)

According to DLT, the relationship between integration cost and the distance between the linguistic entities to be integrated is monotonic. When the distance increases, so does the cost of integrating the current word with the previous word(s). Relating to constituent orderings, DLT suggests that if grammatical alternatives exist, the one that has lower integration cost, or lower syntactic complexity will be preferred. In other words, an optimal constituent ordering will be one that minimizes the distance between words that are syntactically dependent on each other.

In DLT, the distance between a syntactic head and its dependent is measured only on the basis of the number of new discourse referent(s) that come between the two. For example, of the following two sentences, the structure of (3) will be preferred based on DLT.

(3) Dan Jurafsky talked $[_{PP_1}$ with students $]$ $[_{PP_2}$ about why Chinese food has no dessert $]$.

(4) Dan Jurafsky talked $[_{PP_1}$ about why Chinese food has no dessert $]$ $[_{PP_2}$ with students $]$.

For (3), if the preposition is treated as the head within each PP, the cost of integrating the first PP with the head verb of the sentence, *talked*, will be zero. The integration cost

for the second PP will be just one since there is only one NP that is between the head verb and the second PP, *students*. With (4), the integration cost for the first PP is the same. whereas the integration cost for the second PP will be three since there are two new NPs (assuming no previous context before this sentence), *Chinese food* and *dessert* and one new VP headed by *has*. Nevertheless, during the comprehension process, new discourse referents are not the only linguistic items that cause integration cost. In this case when the constituents have an equal number of new discourse referents but are of different length, DLT is not able to make any prediction regarding which ordering might be preferred. For instance, comparing (5) and (6) below, the integration cost for connecting the head verb *looked* and the two dependents *lens* and *sky* is the same, although the lengths of the two PPs are different.

(5) The kid looked $[_{PP_1}$ through the lens $]$ $[_{PP_2}$ at the starry sky $]$.

(6) The kid looked $[_{PP_1}$ at the starry sky $]$ $[_{PP_2}$ through the lens $]$.

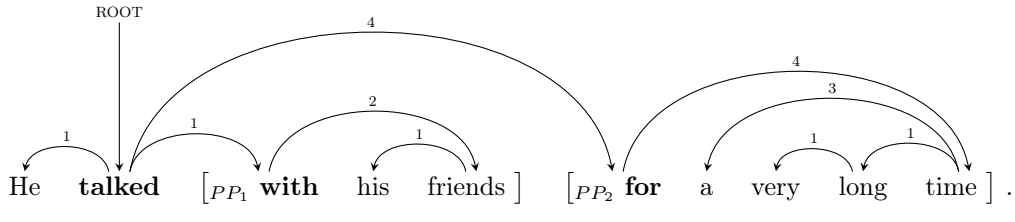
2.2.1.3 Dependency Length Minimization (DLM)

Compared to the aforementioned three principles, DLM has gained in popularity over recent years with the availability of multilingual corpora annotated with morphosyntactic dependency information (see Section 2.1). Recent corpus experiments have proposed DLM as a universal principle of human languages (Futrell et al., 2015a; Liu, 2010) and have shown typologically that grammars tend to minimize the overall or average distance between syntactic heads and their dependents, across 20 languages in Liu (2008) and 37 languages in Futrell et al. (2015a). Other work has tested the predictions of DLM in choosing among multiple possible orders for distinct syntactic structures in specific languages. They have shown that when grammatical alternatives exist for the construction in question, there is a preference for constituents of shorter length to occur closer to their syntactic heads, yielding shorter overall dependency length in the sentence (Jaeger and Norcliffe, 2009; Temperley and Gildea, 2018; Dyer, 2017).

As an illustration of how DLM applies to constituent orderings, let us consider the following two sentences in English. Both (7) and (8) have a VP headed by the verb *talked*. The VPs in the two sentences have the same two PP dependents: *with his friends* and

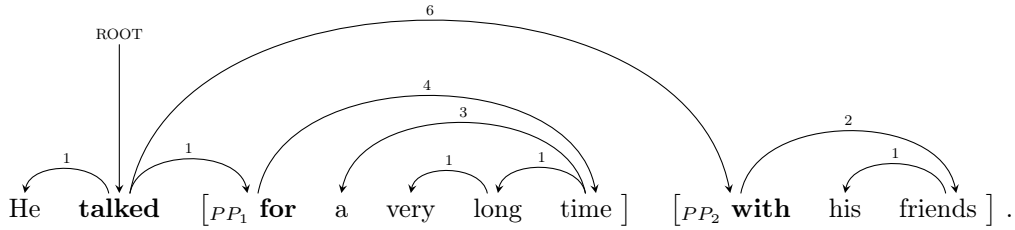
for a very long time.

(7)



total dependency length: $1 + 1 + 2 + 1 + 4 + 4 + 3 + 1 + 1 = 18$

(8)



total dependency length: $1 + 1 + 4 + 3 + 1 + 1 + 6 + 2 + 1 = 20$

As indicated by the syntactic dependency arcs, I consider the two prepositions to be dependents of the verb *talked*. One thing to be pointed out here is that as described in the previous section, the annotation scheme of UD treats the lexical noun within a PP as the head of the PP. By contrast, previous work looking at DLM in double PP orders (Hawkins, 1999; Wiechmann and Lohmann, 2013) has assumed the adposition as the head. Here in particular for ease of demonstration and interpretation for DLM, I opt for function head over content head in the PP and treat the prepositions in both PPs, *with* and *for*, as heads of their respective constituents.[2]

The dependency length between each PP dependent and its head verb is then calculated as the linear distance between the verb and the preposition (the absolute value of the difference in the linear index values of the verb and the preposition within each PP).[3] Comparing (7) and (8), the dependency distance between *talked* and its closest PP is the

[2]Additionally, as a preview, Chapter 3 demonstrates that different choices of the head within a PP do not affect the findings for DLM crosslinguistically.

[3]Following the measure from Hudson (1995), which approximates dependency length as the number of intervening tokens between the head and its dependents, the total dependency length for (7) will be: $0 + 0 + 1 + 0 + 3 + 3 + 2 + 0 + 0 = 9$; the total dependency length for (8) will be: $0 + 0 + 3 + 2 + 0 + 0 + 5 + 1 + 0 = 11$.

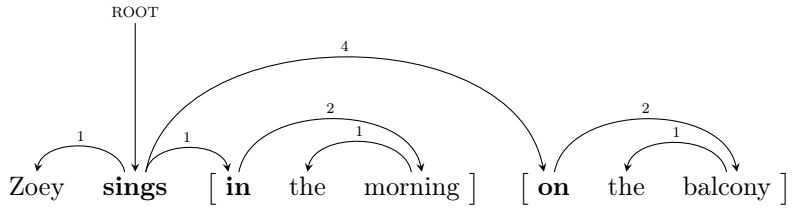
same; whereas the distance between *talked* and the farther PP is shorter in (7). Thus the overall dependency length in (7) is shorter than that in (8), as a result of placing the shorter PP, *with his friends*, closer to its head. On the other hand, If the lexical noun rather than the preposition is treated as the head of the PP, the overall dependency length is still shorter when the short PP occurs closer to the head verb, as in (7). Therefore based on predictions by DLM, the structure of (7) will be preferred.

Various attempts have been made to investigate the role of dependency length in predicting constituent ordering preferences. Though different metrics for measuring dependency length have been proposed, these metrics tend to be highly correlated (Gildea and Jaeger, 2015; Wasow and Arnold, 2003). Most studies have focused on syntactic alternations in English (Gildea and Temperley, 2007; Rajkumar et al., 2016), such as postverbal PP orders (Hawkins, 1999; Wiechmann and Lohmann, 2013), verb particle constructions (Lohse et al., 2004), heavy NP shift (Wasow, 1997a; Arnold et al., 2000) and the dative alternation (Wasow and Arnold, 2003; Bresnan et al., 2007). Other work has examined one or a small number of other languages. For example, psycholinguistic experiments by Yamashita and Chang (2001) showed that speakers of Japanese tend to place long arguments ahead of short ones before the head verb in both transitive and ditransitive construction. Kizach (2012) demonstrated that constituent length affects the ordering of four syntactic structures in Russian: postverbal PPs, the double object structure, adversity impersonals and the order of subject, verb and object. In a corpus analysis, Rasekh-Mahand et al. (2016) modeled the effects of constituent length for relative clause extraposition in Persian. Studies on adjective-noun word order alternations in five Romance languages showed that adjective placement is influenced not only by the dependency distance between the adjective and the noun, but also by the dependency lengths among the dependency subtrees of the NP (Gulordava et al., 2015).

As powerful as its effects are, dependency length on its own is not sufficient for predicting all syntactic orderings across languages. It cannot predict any preferences there might be when switching the order of constituents does not change the overall dependency length of a phrase. As seen in the following examples, both (9) and (10) have two PPs,

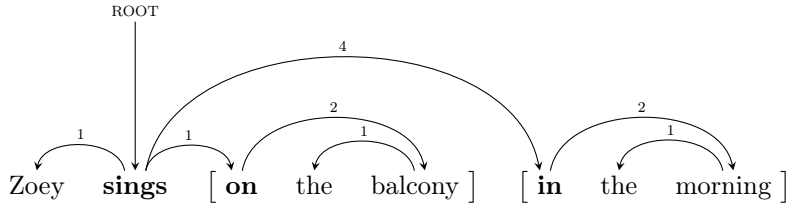
which are of equal length, occurring after the head verb *sings*. Changing the order of the two PPs does not appear to affect the total dependency lengths of both sentences.

(9)



total dependency length: $1 + 1 + 2 + 1 + 4 + 2 + 1 = 12$

(10)



total dependency length: $1 + 1 + 2 + 1 + 4 + 2 + 1 = 12$

2.2.2 Argumenthood status / semantic closeness

The role of argumenthood status in predicting constituent orderings is hardly new. Arguments prefer to be adjacent to their syntactic heads compared to adjuncts, which has been shown extensively for English (Culicover and Jackendoff, 2005; Jackendoff, 1977; Pollard and Sag, 1994), and for other languages as well (Tomlin, 1986; Dyer, 2017), a tendency formulated as the Principle of Argument Closer. Scholars have described argumenthood using different terminologies, such as semantic closeness (Behaghel, 1932) and lexical dependency (Hawkins, 1999). In this dissertation, I use and discuss argumenthood status and semantic closeness interchangeably.

The earliest observation of a correlation between semantic closeness and syntactic proximity has been credited to Behaghel (1932), and is stated as "das geistig eng Zusammengehörige auch eng zusammengestellt wird" (what belongs together semantically is also placed close together) (Wasow and Arnold, 2003). Relating this observation to constituent orderings, Behaghel's principle suggests that constituents that are semantically related should occur closer together syntactically. The preference for semantic closeness preferring syntactic closeness has also been proposed in MiD (see Section 2.2.1.1), which

states that in order to facilitate processing for the human parser, lexical or semantic dependency domains need to be shortened as well. Dyer (2017) suggested that syntactic heads and their dependents can be regarded as being semantically close. If a syntactic head (such as a verb) is semantically dependent on its dependent(s), then the meaning of the head will not be fully recognized or determined until the meaning of its dependent(s) is processed. Thus shortening the distance between the semantically closer dependent with its syntactic head leads to faster meaning construction and efficient processing.

Previous studies have found different ways to define semantic closeness. For instance, in an examination of heavy NP shift, Wasow (1997b) discovered different shifting patterns when the verb and the PP are collocations than when they are not. When the verb and the PP are collocational, in other words, when they are semantically dependent on each other (e.g, *take into account*), the full and appropriate meaning of the verb can only be derived when the PP is processed. Thus there is a greater tendency for the PP to appear closer to the verb, i.e., to shift the NP and to place the PP immediately after the verb. On the other hand, when the verb and the PP are not collocational (e.g., *take to the store*), the proportion of examples where the NP is shifted is much smaller.

Looking at double PP orderings in English (with 394 instances from a mannually collected corpus), Hawkins (1999) noted the significant roles of semantic dependency and syntactic closeness of the PP, namely the PP which is semantically dependent on the head verb tends to be placed closer to the verb. In a similar study, Wiechmann and Lohmann (2013) also found both dependency length and lexical dependency to be strong predictors of PP orderings in English, using 1,256 sentences from the written and spoken sections of the International Corpus of English. Both Hawkins (1999) and Wiechmann and Lohmann (2013) used entailment tests to define the argument status of the PP in relation to the verb. For instance, the PP *on his family* in the sentence *He counts on his family* is an argument of the verb *counts*, since the sentence does not entail *He counts*. By contrast, the PP *in the park* in the sentence *He played in the park* is an adjunct of the verb *played* because the sentence does entail *He played*. With the same entailment tests, Lohse et al. (2004) showed, using several corpora, that the length of the object NP

as well as the argument status of the particle in relation to the verb influence the orders of verb particle constructions in English.

2.2.3 Manner Place Time (MPT)

As proposed in Quirk et al. (1985), the traditional ordering rule for PPs and adverbials in postverbal position in English appears to follow Manner before Place before Time (MPT), as in *Zoey danced [$_{manner}$ elegantly] [$_{place}$ on the dance floor] [$_{time}$ at night]*. By contrast, this rule applies in the opposite direction when the PPs and adverbials occur in preverbal position in verb-final languages, that is, the ordering of preverbal PPs and adverbials follows Time before Place before Manner (TPM) (Hawkins, 1999). While Hawkins (1999) found that MPT plays no significant role in PP ordering in English, Wiechmann and Lohmann (2013) showed that it does have a statistically significant yet weak effect.

2.2.4 Grammatical complexity

According to Wasow and Arnold (2003), the earliest intuition regarding syntactic complexity was given in Chomsky (1975):

> While [...] both [*the detective brought in the suspect*] and [*the detective brought the suspect in*] are grammatical, in general the separability of the preposition is determined by the complexity of the NP object. Thus we could scarcely have [...] *the detective brought the man who was accused of having stolen the automobile in.*

Though the notion of *complexity* is too vague, previous research has proposed various ways of quantifying the level of complexity in different syntactic constructions. For instance, in Wasow and Arnold (2003), between two phrases of equal length, the one with an embedded clausal structure is considered more complex. To test whether grammatical complexity and dependency length are two separate factors, they conducted both a questionnaire study and a corpus analysis investigating whether complexity predicts postverbal constituent ordering in certain English constructions, when the length difference between the sentences is controlled for. Their results showed positive effects for grammatical complexity. For example, complex NPs are more likely to be shifted than simple NPs in heavy

NP shift. Though both simple and complex NPs tend to appear after the particle in verb particle constructions, the verb and the particle are less likely to be split when the NP is complex. Similar results were given for the dative alternation. More complex NPs are preferred in the structural final position, regardless of whether it is the complexity of the direct object or the indirect object that varies. When complexity and dependency length both vary, their results indicated that the predictive accuracy increases when the two factors are combined. Wasow and Arnold (2003) suggested that these observations could be explained in terms of production and comprehension efficiency. Planning or parsing constituents with relatively long length or complex structures is more likely to incur more memory and resource allocations.

The syntactic complexity of an NP in English has been measured in more fine-grained ways in later work, such as the number of phrasal nodes or total nodes within the NP (Berlage, 2014), whether the NP contains prenominal and/or postnominal modifiers, the phrasal type of the modifiers of the NP (e.g. adjective, PP, relative clause) as well as the length of the NP (Schilk and Schaub, 2016). Above the NP level, in Rajkumar et al. (2016), factors such as dependency length, average surprisal and embedding depth were also included as approximates for the syntactic complexity of a sentence, and these constraints were employed to predict a range of syntactic alternations in English.

2.2.5 Discourse structure

The most important principle that characterizes the role of discourse structure in constituent orderings is *given before new*, which states that *given* or *old* information previously mentioned in the context or discourse tends to appear before relatively *new* information. The earliest proposal for this principle was perhaps by Behaghel (1932): “es stehen die alten Begriffe vor den neuen" (old concepts come before new ones). This generalization was then later captured as “given before new" by Gundel (1988). The effects of discourse information status have been empirically attested in various context. For example, the production experiments of Bock and Irwin (1980) showed that given information tends to be produced earlier. Similar results have been replicated in Ferreira and Yoshita (2003) with Japanese. Arnold et al. (2000) demonstrated that in both heavy NP shift and the

dative alternation in English, speakers prefer to produce the relatively newer constituent later. In a comparative study of English and Spanish, Prat-Sala and Branigan (2000) also presented evidence that the difference between the accessibility of the agent and that of the patient, or their relative information saliency, affects whether speakers will opt for the active or the passive construction.

2.2.6 Verb bias

Numerous studies have demonstrated that language users are sensitive to the probabilistic information of verb subcategorization frames. Specific verbs have biases towards appearing in a particular syntactic alternation. The probability of different subcategorization frames for the verb can predict processing preferences as well as argument structure realizations. Most studies along these lines have focused on aspects of ambiguity resolution in language comprehension. The general finding is that if the construction conforms to the structural bias of the verb, it will facilitate processing (Clifton et al., 1984; Hale, 2001; MacDonald, 1994; Trueswell, 1996; MacDonald et al., 1994; Garnsey et al., 1997; Trueswell, 1996; Trueswell et al., 1993). For example, the verb *suggest* has a stronger preference for taking a sentential complement (e.g. *I* ***suggest*** *that it is time to investigate verb bias.*) than a direct object (e.g. *The reviewers* ***suggest*** *more examples.*) (Gahl and Garnsey, 2004). Accordingly, a sentence where *suggest* is followed by a sentential complement is easier to process compared to one where *suggest* takes a direct object. Other experiments have shown the effects of verb bias on structural priming. The stronger bias a verb has for a particular subcategorization frame, the stronger the priming effect that is observed (Gries, 2005; Peter et al., 2015).

By comparison, the role of verb bias in language production and constituent orderings has not been subjected to such extensive investigation as it has been in comprehension experiments. A number of corpus studies looking at the dative alternation and heavy NP shift in English (Wasow and Arnold, 2003; Wasow, 1997a,b) have shown that the likelihood of a verb appearing in the double object structure or the prepositional object structure is indicative of the orderings observed. In a series of production studies, Stallings et al. (1998) demonstrated that speakers tend to use heavy NP shift with a verb like *explain*

in comparison to a verb such as *release*. This is because *explain* more regularly takes a sentential complement that is not typically an NP, which prefers to be adjacent to the verb, leading the NP to be shifted.

Why would verb idiosyncracy matter in the first place? To explain this from a psycholinguistic perspective, Wasow (1997a,b) proposed the concept of early versus late commitment, which aims to potentially separate speaker-based and listener-based accounts of language processing. From the listeners' perspective, it is better if the upcoming syntactic structure can be predicted, or recognized earlier, which lessens memory load for the parser. In other words, listeners are more likely to prefer early commitment to the full syntactic structure. Whereas for speakers, postponing the decision for the whole syntactic structure allows more time and scope for how the sentence might continue and is good for utterance planning and production. In this way speakers will have more time to prepare, formulate and articulate what they intend to say, and therefore they would prefer late commitment.

Wasow (1997a) examined this proposal with instances of heavy NP shift. He compared the proportion of heavy NP shift for both transitive and prepositional verbs. Since transitive verbs require an NP object, if the structure places the NP right next to the head verb, when the listeners get to the NP, they have no extra information regarding how the sentence might continue, as there is no way to infer whether there would be a PP structure in the following. On the other hand, if the PP is ordered before the NP, when the listeners hear the utterance of *to* (the function head of the PP in the instances examined), they will realize there is an NP to come. Therefore the early commitment regarding how the sentence is structured will be fulfilled.

However, for prepositional verbs, things are different. If a prepositional verb is followed immediately by an NP object, listeners will know to expect a PP afterwards. Whereas if a PP is placed adjacent to the verb, the sentence can continue in different ways. For instance, the sentence might end just after the PP, or take on an additional NP, and so on. For listeners, if the NP is not much heavier, placing the PP right after the verb makes parsing relatively more difficult. They will prefer a V-NP-PP construction for the purpose

of early commitment. By contrast, transitive verbs will have a larger number of heavy NP shift cases than prepositional verbs.

The opposite predictions can be made for speakers, who are more likely to prefer late commitment of the entire syntactic structure. After producing a transitive verb, there will be less taxing and demand on memory load if an NP object can be uttered immediately, which leaves more planning time for speakers. Similarly, placing a PP adjacent to a prepositional verb will keep options open for how the full utterance will turn out to be. From this perspective, prepositional verbs will be more likely to have heavy NP shift than transitive verbs.

Chapter 3

Mixed Evidence for Crosslinguistic Dependency Length Minimization

3.1 Introduction

This chapter investigates the crosslinguistic extent of DLM (see Section 2.2.1.3). Specifically, this chapter asks two questions. First, are shorter dependencies preferred in syntactic constructions with flexible constituent orderings across languages? It is currently unclear to what extent there is a typological preference for DLM in sentences with syntactic alternations, when language users presumably have a choice between different orders of the constituents.

Secondly, how strongly does DLM apply in defining the ordering preferences of languages with different syntactic characteristics? Very few studies in the literature have tried to address this issue (Futrell et al., 2015a; Gildea and Temperley, 2010; Liu, 2010; Ros, 2018). In the corpus experiment of Gildea and Temperley (2010), the grammar of German appeared to minimize dependency length to a lesser extent than that of English. Results from Futrell et al. (2015a) suggested that head-final languages such as Japanese and Turkish tolerate longer dependencies than head-initial languages like Italian and Indonesian, though without explicit statistical tests. Liu (2008) demonstrated that among the 20 languages examined, Mandarin Chinese has the longest mean dependency length. Looking at transitive and ditransitive constructions in Basque, Polish and Spanish, Ros (2018) showed that Basque, an OV language with a mix of verb-medial and verb-final

orders, has weaker DLM effects compared to the other two languages.

I examined these two questions using PP typology as a test case. I focused specifically on sentences with VPs that have exactly two PP dependents appearing on the same side of the head verb, the ordering of which permits flexibility in some contexts (e.g. *She danced with the band at the dinner party* vs. *She danced at the dinner party with the band*). These cases allow language users to presumably have a choice regarding the order of the two PPs .

Going beyond previous studies on syntactic ordering preferences, including experiments on PP order which have mainly looked at English and used relatively small amounts of data (Hawkins, 1999; Wiechmann and Lohmann, 2013), I leveraged multilingual corpora data for 34 languages from the Universal Dependencies project version 2.5 (UD) (Zeman et al., 2019). I asked:

(1) whether there is a cross-linguistic tendency for the shorter PP to appear closer to the head verb;

(2) how the strength of DLM varies in different PP ordering patterns across languages with different word order features.

3.2 Data and methods

3.2.1 Data and preprocessing

In the present context I focused on contemporary languages and their treebank data from UD. I searched in every treebank for sentences with VPs containing exactly two PP dependents that both occur on the same side of the head verb. Abiding by the UD annotation scheme, I selected verbs with a POS tag of VERB, which denotes only main verb or content verb. PPs with no lexical head, in other words, PPs that only consisted of a preposition or a postposition were not considered. The dependency relation between the lexical head of each PP and the head verb in the instances that I extracted was always *obl* (oblique). The POS of the function head within each PP was ADP (adposition) and its dependency relation with the lexical head was always *case*, which is "used for any case-marking element which is treated as a separate syntactic word (including

prepositions, postpositions, and clitic case markers)".[1] Following The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), I coded the word order features as well as the language family and genus of each language. In cases where there were multiple treebanks for the same language, I first calculated the effects of dependency length for each treebank individually. Since I found DLM differences between the treebanks of the same language to be very small, I combined them into one for my analysis. Languages with fewer than a total of 50 sentences that fit the search criteria (i.e. containing a VP with two PPs attached to the same side of the verb) were not included in the analysis, yielding a dataset for 34 languages. These languages ended up being mostly Indo-European, with the exceptions of Arabic and Hebrew (Afro-Asiatic), Indonesian (Austronesian), Japanese (Japanese), and (Mandarin) Chinese (Sino-Tibetan).[2]

3.2.2 Measures of dependency length effects

To estimate the effect of dependency length on PP ordering, I measured the length of the PP closer to the verb and of the PP farther from the verb as the number of tokens in each PP. I approximated phrase length using the number of tokens according to the treebank tokenization. If the preference for DLM holds for PP orderings, I would expect the proportion for when the shorter PP is closer to be significantly higher than that for when the longer PP is closer in every language.

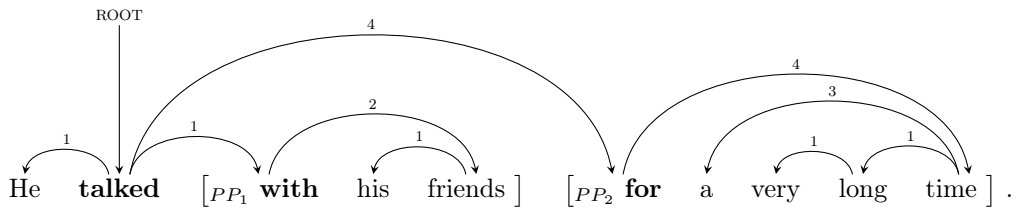
Recall the example in Section 2.2.1.3, shown below, where the dependency length between the head verb and each PP is measured as the linear distance from the verb to the function head of each PP (the absolute value of the difference in the linear index values of the verb and the adposition within each PP). If (1) is the original instance observed in the corpus, while (2) is the structural variant where the order of the two PP is switched, The length of the PP farther away from the head verb (which is 5) minus the length of the PP closer to the verb (which is 3) is conceptually as well as mathematically equivalent to the overall dependency length of the structural variant minus the overall dependency length of the original instance. In other words, the measure I described above is the same

[1] `https://universaldependencies.org/u/dep/case.html`

[2] UD 2.5 has treebanks for both traditional and simplified Chinese. I focused only on simplified Chinese.

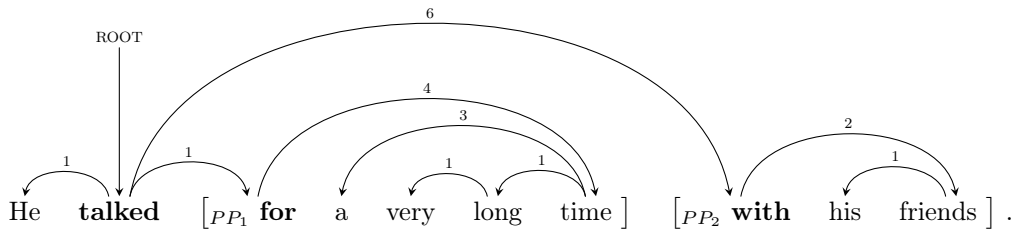
as comparing overall dependency length when choosing the function head of each PP.

(1)



total dependency length: $1 + 1 + 2 + 1 + 4 + 4 + 3 + 1 + 1 = 18$

(2)



total dependency length: $1 + 1 + 4 + 3 + 1 + 1 + 6 + 2 + 1 = 20$

For hypothesis testing of the effect for dependency length and estimates of confidence intervals, I employed bootstrapping (Efron, 1979), a resampling method. The basic procedure for the computation is as follows. (1) From a dataset of n VP instances for a certain language after preprocessing, I drew a sample of n VP instances with replacement, where some of the cases from the original dataset could appear multiple times; (2) given this sample, I calculated the proportion for three different cases: when the shorter PP occurs closer, when the longer PP is closer, and when the two PPs are of equal length; (3) I repeated steps (1) and (2) for 1,000,000 iterations to derive an empirical resampling distribution for each of the three different case scenarios described in (2); (4) for each resampling distribution, I calculated the mean and 95% confidence interval.

As described in Section 2.1, the annotation scheme of UD favors content head over function head for comparable representations of syntactic constructions across languages. To evaluate whether this particular annotation choice would affect the results, I also measured dependency length as the linear distance between the verb and the content head within each PP, repeating the same statistical procedures above. There was no observed difference (see Section 3.3) in the findings compared to when choosing the function head.

3.3 Results

Overall the dataset shows three different PP ordering patterns:

- one for languages with head-initial PPs after the head verb (e.g. Indonesian);
- one for languages with mixed PP orders where head-initial PPs can appear after or before the head verb (e.g. German);
- one for languages with head-final PPs before the head verb (e.g. Japanese).

Taking a data-driven approach, I grouped the languages based on their PP ordering patterns first, while taking into consideration the grammatical and syntactic features of different language types as well within my analysis.

3.3.1 Languages with head-initial PPs after head verb

For twelve languages in the dataset, when the head verb has two PP dependents occurring on the same side, they appear as head-initial PPs after the head verb. These languages are all head-initial, or have a dominant VO order. The prediction made by DLM is that the PP that is the first in the sequence of the two PPs, and which is closer to the head verb, should be preferably shorter. As shown in Figure 3.1, green bars represent the proportion of cases when shorter PPs are closer, grey bars the cases when longer PPs are closer, and yellow bars the cases when the two PPs have equal length. Of the twelve languages, a full eleven show a strong preference for DLM, except for Latvian. And on average, the number of sentences with the shorter PP closer to the verb is 3.9 times that of sentences with the longer PP closer to the verb. None of these languages displays a preference for the longer PP to occur first.

3.3.2 Languages with head-initial PPs after or before head verb

Sixteen languages in the dataset, shown in Figure 3.2, exhibit mixed PP ordering patterns. These languages fall into three different subtypes (genera): Germanic, Slavic and Romance, most of which are mixed-type languages. For all sixteen languages, I observe the same PP orders as those in Figure 3.1: when a head verb has two PP dependents on the same side, they occur as head-initial PPs after the head verb. In addition, they also

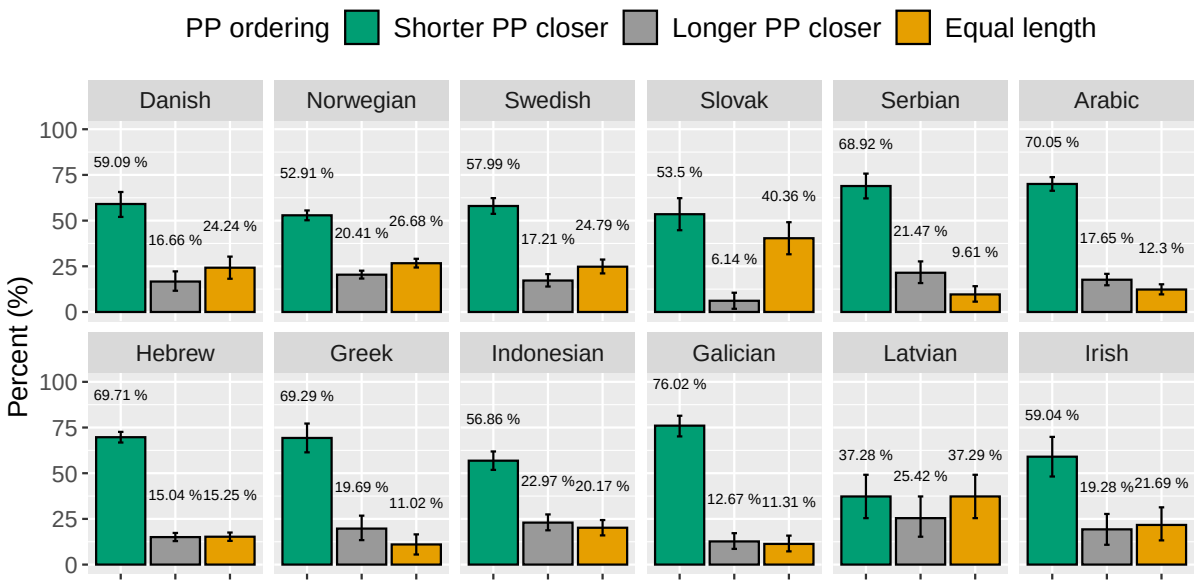


Figure 3.1: DLM in languages with head-initial PPs after the head verb. Error bars represent 95% confidence intervals.

have cases in which two head-initial PPs occur before the head verb. As an illustration of this ordering pattern, consider the following sentences in German.[3] If dependency length exerts a positive effect, the structure of (3) will be preferred to that of (4) since in (3) the shorter PP is closer to the head verb. In this case the shorter PP is actually the second PP in the sequence of the two PPs.

(3) dass er [**an** dem Tisch] [**mit** ihr] **isst**

that he [at the table] [with her] eats

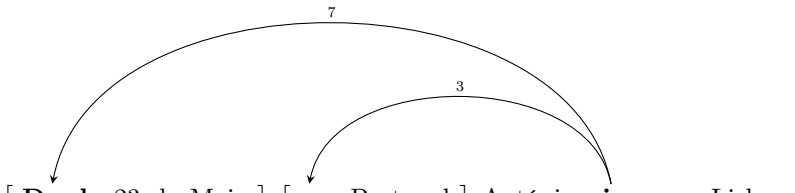
(4) dass er [**mit** ihr] [**an** dem Tisch] **isst**

that he [with her] [at the table] eats

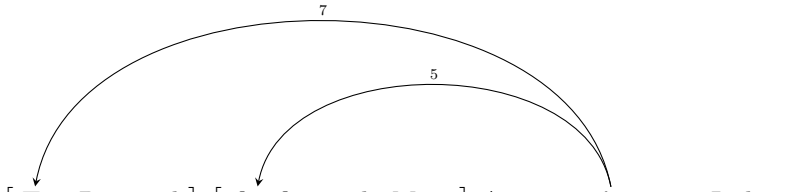
[3] Again for ease of demonstration and interpretability, here I illustrate examples opting for the function head within each PP. If choosing the content head, the overall dependency length between the verb and its two PP dependents is 4 in (3) and 5 in (4), the former still being shorter than the latter.

'that he eats with her at the table.'

As shown in Figure 3.2a, when the PPs occur after the head verb, dependency length appears to be a strong predictor of PP ordering, and the results are comparable to the results in Figure 3.1. There is a dominant preference for the shorter PP to appear closer to the head verb, and to occur first, and none of these languages shows a tendency to put the longer PP closer to the head verb. The average ratio between the number of VPs with the shorter PP closer to the verb and the number of VPs with the longer PP closer is 3.0. On the other hand, when the PPs occur before the head verb as in Figure 3.2b, any preference for DLM almost disappears across the sixteen languages. Though in Dutch there exhibits a tendency to put the shorter PP closer to the head verb when the two PPs occur preverbally, the extent is less pronounced compared to when the two PPs appear postverbally in the language (1.4 vs. 2.2). Most of the Romance languages here in particular, except Portuguese and Romanian, actually demonstrate a robust short-before-long ordering preference, where the shorter PPs tend to appear first, and farther away from the head verb.

(5) 

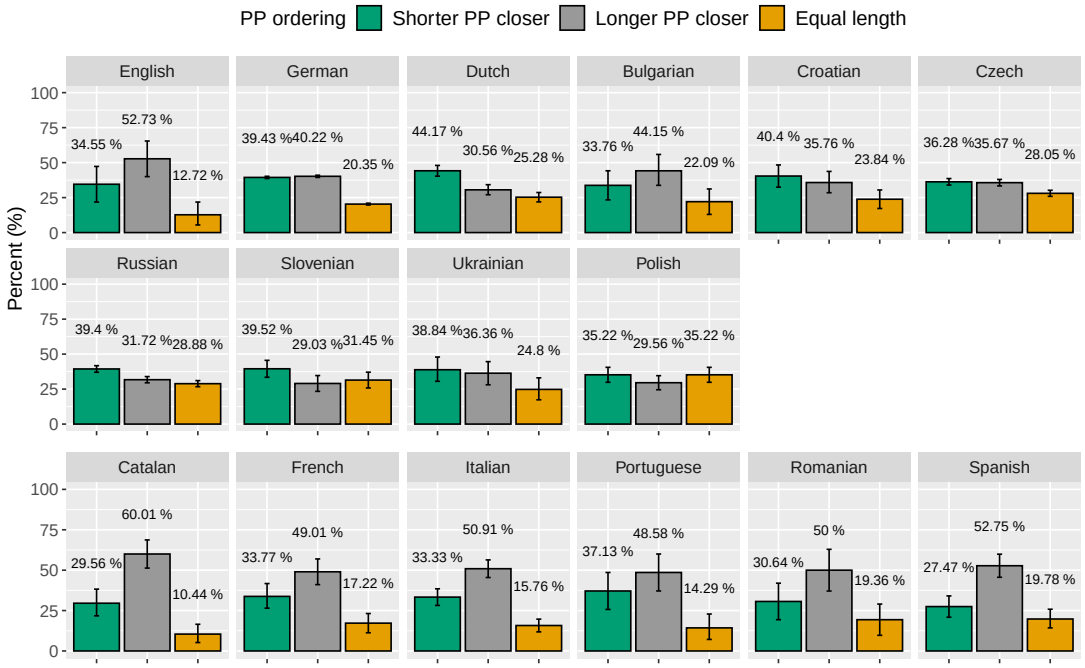
[Since 23 of May] [in Portugal] António lived in Lisbon

(6) 

[In Portugal] [since 23 of May] António lived in Lisbon

'Since May 23, in Portugal, António lived in Lisbon.'

(a) DLM in mixed-type languages when PPs appear after the verb.



(b) DLM in mixed-type languages when PPs appear before the verb.

Figure 3.2: DLM in mixed-type languages in which PPs can appear after or before the verb. Error bars represent 95% confidence intervals.

It is worth noting that there are both different and similar syntactic characteristics

among the three language subtypes. Within the three Germanic languages, English differs from German and Dutch, in that it is dominantly SVO. Though English has relatively rigid word orders, in certain pragmatic or social context, the object is preposed before the head verb, forming an OSV order (Prince, 1988). German and Dutch are both classified as non-rigid OV, on the other hand, allowing a mixture of verb-medial and verb-final clausal structures. For these two languages, in main clauses with one or more auxiliary verbs as well as in subordinate clauses, the finite and lexical verbs regularly come in the clause-final position after the object. When the main clauses do not have an auxiliary verb, the finite and lexical verbs move to the front of the object (Hawkins, 2019). Despite the relatively rigid requirement to place the verb in the final position given the aforementioned conditions, both German and Dutch do allow flexibility in the ordering of subject, direct and indirect object (MacWhinney et al., 1984; Kempen and Harbusch, 2004). Both the Slavic and Romance languages have predominantly VO constructions. In contrast to German and Dutch, they do not exhibit different verb positions or verb movements that depend on clause type and on whether the clause has an auxiliary verb. Nevertheless, it has been shown that Slavic and Romance languages also allow variation in the ordering of core arguments of the head verb (Siewierska, 1998; Liu, 2010).

Despite these partially similar and distinct syntactic properties of the Germanic, Slavic and Romance languages, if I combine the observations in both Figure 3.2a and Figure 3.2b, it is obvious that the contrast regarding the tendency for DLM between postverbal and preverbal PP orderings is shared across the three language subtypes: there is a strong preference for shorter dependencies in postverbal domains, whereas the preference is much weaker or disappears in preverbal structures. Though this pattern was theoretically predicted based on the EIC principle and the phrase structural framework of Hawkins (1994), he did not have comparative preverbal and postverbal data on which to test this. To the best of my knowledge, the results here are the first empirical demonstration of this clean and sharp distinction.

A puzzle remains, which is why there are instances with two head-initial PPs before the head verb in these languages in the first place, especially in Romance and Slavic. A recent

study by Futrell et al. (2015b) has quantified the degree of word order freedom across a variety of languages. Their results show that non-English Germanic, Slavic and Romance languages have more flexible orderings compared to other language genera. Thus one reason might be due to the relative flexibility of certain types of PPs in relation to the head verb, especially those that denote temporal expressions or locations. For instance, the PP *en casa* in Spanish, which describes the location *at home*, is also able to appear either before or after the head verb. Moreover, it is possible that the freedom of these PPs is less dependent on the rigidity of a language's full word order profile. This explains why I also see the same variable ordering pattern in English, though it has a much smaller number of total occurrences (55). These types of PPs are likely to appear as the one that is farther away from the head verb in the sequence of two PPs, and they might tend to be relatively short on average as well. In a qualitative examination of all 55 instances in English which had two head-initial PPs before the head verb, among the 29 cases which had a short-before-long orderings, 12 had a temporal PP occurring first (41.4%).

Besides the sixteen languages seen above, three other typologically mixed languages, Afrikaans, Persian and Chinese, present only instances where two head-initial PPs occur before the head verb. Among these three, Afrikaans and Persian are non-rigid OV languages and they are both prepositional. Afrikaans shows a significant preference for the shorter PP to appear closer to the head verb, though the tendency for DLM is weaker than in cases which have head-initial PPs after the head verb. The average ratio between the number of VPs with shorter PPs closer and the number of VPs with longer PPs closer is 2.2, compared to 3.9 for languages in Figure 3.1 and 3.0 for languages in Figure 3.2a. In contrast to German and Dutch in Figure 3.2, it seems surprising that there are no VP instances with head-initial PPs placed after the head verb in Afrikaans. This might result from the specific genre of the Afrikaans corpus, which is mainly government texts. Additionally, the number of Afrikaans tokens in the dataset is small (92). On the other hand, Persian only exhibits a weak effect for dependency length, while in Chinese I observe no preference for DLM at all.

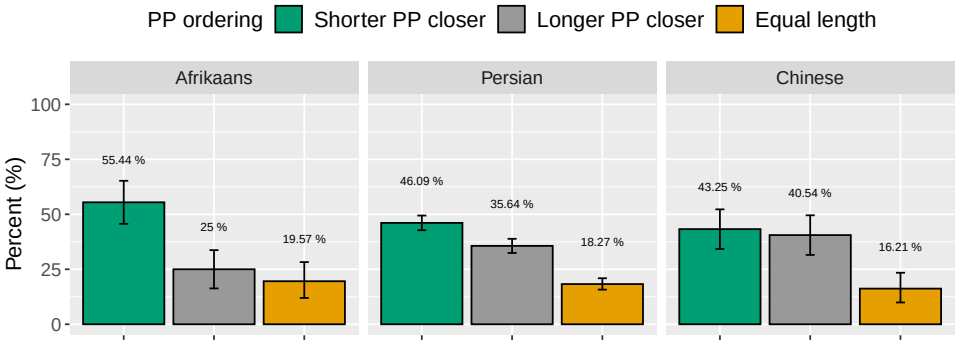
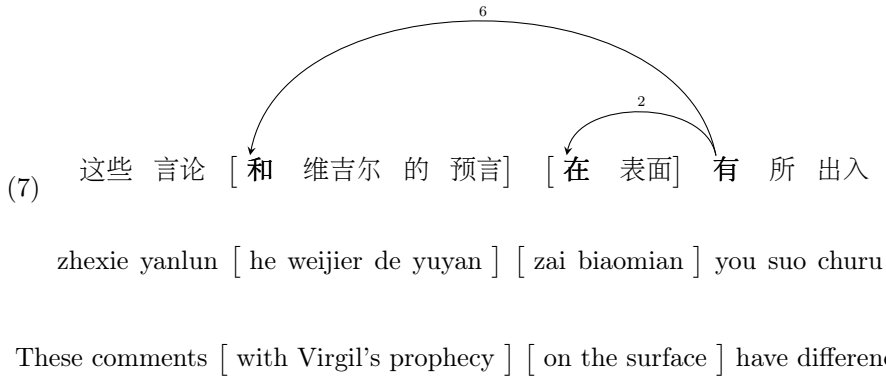


Figure 3.3: DLM in Afrikaans, Persian and Chinese. Error bars represent 95% confidence intervals.

One unique characteristic about Chinese that should be pointed out is that in contrast to the other languages studied here, Chinese has profound inconsistencies in the headedness of its various structures (Dryer, 1991). Although the Chinese VP instances extracted all have two head-initial PPs before the head verb, as illustrated below, the adposition system of Chinese has been argued to be more complicated (Li and Thompson, 1974). While previous work has quantitatively classified Chinese as an SVO language (Sun and Givón, 1985; Mei, 1980; Liu et al., 2009), it has been suggested that Chinese has both prepositions and postpositions that form head-initial (e.g. (9)) or head-final (e.g. (10)) PPs respectively (Hawkins, 1983; Wang and Sun, 2015). This is different from the languages I have discussed so far, which are all prepositional (languages in Figure 3.1, Figure 3.2, Figure 3.3).



(8) 这些 言论 [在 表面] [和 维吉尔 的 预言] 有 所 出入

zhexie yanlun [zai biaomian] [he weijier de yuyan] you suo churu

These comments [on the surface] [with Virgil's prophecy] have differences.

'These comments have differences on the surface with Virgil's prophecy.'

(9) 在 美国
in America
'in America'

(10) 书包 里
bag in
'in the bag'

(11) 从 法庭 上
from court on
'from court'

Nevertheless, cases such as (11) also exist in Chinese, in which a phrase takes both a preposition and a postposition. Cai (2013) noted that the preposition and postposition in such contexts together form a circumposition, whereas Zhang (2013) proposed that what appears in the post-nominal position should be treated as a bound localizer rather than a postposition.

In the Chinese data, when examples similar to (11) occur, the POS of the post-nominal form is annotated as an adposition, and it is treated as a clausal modifier of its head noun. To test if this particular annotation scheme affects DLM, I removed sentences with cases as (11). After the removal there was 78 sentences in total, with 42.31% having the shorter PP closer to the head verb and 35.9% having the longer PP closer to its head.

3.3.3 Languages with head-final PPs before head verb

In the dataset, three languages turned out to be rigid OV, which are Japanese and two Indic languages, Hindi and Urdu. For these languages, when a head verb has two PP dependents, they appear as head-final PPs before the head verb. If DLM holds, when the shorter PP occurs closer to the head verb, it is also the second PP in the sequence of the two PPs. This is similar to instances with head-initial PPs before the head verb, and contrasts cases with head-initial PPs after the head verb. As illustrated in the examples below, the ordering structure of (12) will be preferred to that of (13), because by putting the shorter PP, 自分 で, second and closer to the head verb, the overall dependency distance is shorter in (12).

(12) 彼 は ［ 赤い ペン で ］ ［ 自分 で ］ 書いた

He ［ a red pen with ］ ［ himself by ］ wrote.

(13) 彼 は ［ 自分 で ］ ［ 赤い ペン で ］ 書いた

He ［ himself by ］ ［ a red pen with ］ wrote.

'He wrote by himself with a red pen.'

As presented in Figure 3.4, the preference for DLM in the three rigid OV languages is much weaker overall. There appears to be only a mild effect of dependency length for Japanese. The ratio between the number of VPs with the shorter PP closer to the verb and the number of VPs with the longer PP closer is 1.5. Though this is substantially lower than the 3.9 and 3.0 average ratio, respectively, for the languages in Figure 3.1 and those in Figure 3.2a. The contrast between Hindi and Urdu and languages with postverbal head-initial PPs is even more pronounced, with there being no observed tendency for shorter dependencies at all.

The observations presented here go against findings for rigid head-final languages in previous studies (e.g. Japanese (Yamashita, 2002); Korean (Choi, 2007)). Their experiments examined the relative order of subject, direct and indirect object in relation to the head verb, and their results showed robust preferences for DLM. By contrast, I use PP obliques as a test case, the ordering of which allows more word order flexibility and is less governed by specific grammatical constraints, compared to the order of the three core arguments as mentioned above. This offers more possibilities for probing the direct role of dependency length as a predictor for constituent orders and casts some doubt on its overall effects.

Figure 3.4: DLM in languages with head-final PPs before the head verb. Error bars represent 95% confidence intervals.

I understand that the sample size of rigid OV languages in the dataset is quite small in contrast to the dominantly VO and mixed-type languages previously seen. Accordingly, I am cautious about drawing any conclusive statements here regarding the different preference for DLM between the different language (sub)types. More work is needed in the future to test whether the tendency for shorter dependencies exists and is consistent in other rigid head-final languages. Nevertheless, this does not necessarily undermine my findings here, as I believe the reasons for the small sample size presented in Figure 3.4 are not arbitrary.

First, unlike prepositions in head-initial languages, postpositions in head-final languages are relatively less “productive" (Tsunoda et al., 1995; Hawkins, 2014), due to the fact that they may have become lost in language change or may have developed into

suffixes rather than staying as freestanding grammatical elements. This has possibly led head-final languages to have smaller numbers of different types of postpositions compared to the numbers of different types of prepositions in languages like English. To validate this, I extracted all PP obliques that are dependents of verbs from the English treebank, then counted the different types of prepositions that are the functional heads of these PPs. I did the same for the different types of postpositions in the treebanks for Japanese, Hindi and Urdu as well. As shown in Table 3.1, the three rigid OV languages have smaller inventories of postposition types in comparison to the prepositions in English.

	English	**Japanese**	**Hindi**	**Urdu**
preposition	158			
postposition		68	123	112

Table 3.1: Counts of preposition or postposition types in English, Japanese, Hindi and Urdu.

These numbers lead to my second point, which is that what is expressed as a PP oblique in English is possibly realized in a different syntactic structure in these head-final languages. For example, dative verbs in English such as *give* and *send* can have two alternative argument structures, the double object structure, as in *I gave her a book*, as well as the prepositional dative structure, as in *I gave a book to her*. In the latter the indirect object *her* is marked by the preposition *to* and the two together form a PP oblique. Whereas in Japanese, as shown in (14), the indirect object for the corresponding verb of *give* is always realized as an NP via the attachment of the case marker に . The case markers are different from postpositions or prepositions as observed in PPs, which potentially results in fewer instances where the head verb has two PP oblique dependents on the same side.

(14) 私 は　彼女 に　本 を　与えました
I -NOM she -DAT book -ACC gave
'I gave her a book.'

3.4 Discussion

3.4.1 Crosslinguistic effects and the efficacy of DLM

Now I turn back to my initial questions. First, is there a preference for DLM in syntactic constructions with flexible constituent orderings across languages? I show empirically that the answer is yes. Dependency length is an effective typological determinant in PP order typology. Overall there is a strong crosslinguistic tendency for shorter PPs to appear closer to the head verb. If the preference for shorter dependencies is driven by processing ease (Gibson, 1998; Hawkins, 2015; Tily, 2010; Gibson and Wu, 2013), the fact that PP orders with DLM are more prevalent across languages suggests the ordering structure that is relatively easier to process is more preferred. This contributes to the research program of processing typology (Hawkins, 2007) which asserts that the distribution of various syntactic variations is shaped by pressure associated with language usage (Bybee, 2006; Ellis, 2002; Jaeger and Tily, 2011), and that language universals can be explained in terms of facilitating processing efficiency and robust communication (for a review, see Gibson et al. (2019)).

The second question I raised at the beginning is how the extent of DLM in ordering preferences varies for languages with different syntactic characteristics. Overall, I have shown that the efficacy of DLM depends on the specific ordering structures in different language types, yielding two general patterns. First, the preference for shorter dependencies is the most pronounced when head-initial PPs appear after the verb. Whereas when they occur before the head verb, there is a much weaker or no tendency for DLM. This contrast is the most visible in languages with mixed PP orders, shared across different language genera for most of the mixed-type languages in the dataset. Secondly, acknowledging their limited sample size in the dataset, I also find no or a much weaker preference for DLM in rigid head-final languages, which have head-final PPs before the head verb. Taking the two general patterns into consideration, the results indicate that across languages, preverbal PP orderings have no pronounced tendency for DLM in comparison to postverbal PP orders. This observation appears to be less dependent on the structural characteristics of specific language types and seems to hold regardless of whether the PP

is prepositional or postpositional.

Based on previous work, it appears that two considerations could explain the weaker tendency for DLM in certain language types: word order freedom and rich case marking. These two properties have been shown to be strongly correlated across a number of languages (Spair, 1921; Futrell et al., 2015b). Gildea and Temperley (2010) attributed the reason for why German has longer dependency length than English to the fact that German has comparatively more word order freedom and more verb-final structures. This might lead to more inconsistent head-dependent orderings and longer preverbal dependencies. Both Ros (2018) and Futrell et al. (2015a) suggested that head-final languages exhibit longer dependencies due to their rich case marking systems, which lead to ordering flexibility.

However, word order freedom and rich case marking are not always correlated across languages. When a language has a rich case marking system, that does not necessarily mean it has flexible word orders (e.g. Icelandic (Kiparsky, 1996)). On the other hand, languages such as Lau (Enfield, 2009) and Riau Indonesian (Gil, 2017) have relatively free word orders, but they do not have rich case marking. Accordingly, how to precisely tease apart the separate roles of the two aspects in constituent ordering preference and how they are related to the findings here need further quantitative validations.

As analyzed in Section 3.3.2, the relative flexibility of certain types of PPs could explain why there are instances with two head-initial PP dependents before the head verb in the first place. Though it still remains unclear whether the structure would exploit this freedom and put the shorter PP closer to the head verb. It is possible that PPs which can only appear either preverbally or postverbally have more fixed positions in general, and might prefer ordering proximity to the head verb, regardless of whether that leads to overall shorter dependencies. In addition, most of the mixed-type and the rigid OV languages in my dataset have rich case marking. Both types also have longer dependencies in preverbal PP orderings. Nevertheless, when a language or a structure has more case markers but no pronounced tendency for DLM, this does not mean that the long dependencies are caused by the case markers.

3.4.2 Speakers' vs. listeners' perspective

One possible reason for some of the variants in the results may lie in processing differences between syntactic comprehension and production, that is, whether the ordering patterns are mainly for the benefit of the listener or the speaker. In addition, languages with distinct word order features favor different factors or adopt language-specific strategies in order to facilitate either comprehension or production ease (Ueno and Polinsky, 2009).

From the listeners' perspective, putting shorter constituents closer to their heads shortens the overall dependency lengths of the sentence, and will potentially ease comprehension and online parsing. This has been widely documented in comprehension tasks for different structures in a number of languages (Gibson, 1998, 2000; Gibson and Wu, 2013; Levy et al., 2013; Liu, 2008), where longer dependency lengths lead to more processing difficulty, measured by reading times. Similarly here, the PP orderings might place the shorter PP closer to the head verb for the benefit of syntactic comprehension.

However, several studies examining the comprehension of verb-final structures indicate that the comprehension process for these structures is actually faster when the opposite patterns to DLM are observed (Levy, 2013). For instance, it has been shown that in certain verb-final constructions in German, adding preverbal dependents between the arguments and the clause-final verb makes the sentence easier to process, at the processing of the verb at least (Konieczny, 2000; Konieczny and Döring, 2003). Vasishth and Lewis (2006) demonstrated that interposing dependents between arguments of the clausal-final verb facilitates processing in Hindi. Evidence from these studies suggests that adding more dependents before the verb makes the verb more predictable, and faster to read. The connection between these studies and mine may not be seemingly obvious, as all the VP instances here have exactly two PP dependents. Nevertheless, it is possible that the absence of DLM for certain PP orderings can be explained along similar lines. When the two PPs both appear before the head verb, the head verb might have more predictability if the PPs are ordered as short-before-long; or the longer PP might be more predictive of the following head verb.

From the speakers' perspective, however, as syntactic production is an incremental

process (Ferreira and Swets, 2002), the speakers might not abide by DLM when ordering their utterances. They might opt for constituent orderings that will facilitate utterance planning and ease production efforts (Arnold et al., 2000). In this case speakers are likely to produce short constituents first which are formulated faster (De Smedt, 1994; MacDonald, 2013), more predictable given the preceding context or easier to access conceptually (Lohmann and Takada, 2014), and choose to postpone the longer ones in order to have more planning time for the whole sentence they intend to speak. This is shown consistently here across languages when they have head-initial PPs after the head verb. In cases where the two PPs both occur before the same head verb, regardless of the headedness of the PP, there might also be a tendency for the shorter PP to appear before the longer one.

Nevertheless, this short-before-long preference in constituent utterances has been countered by production studies of head-final languages such as Japanese (Yamashita and Chang, 2001) and Korean (Choi, 2007). As discussed in Yamashita and Chang (2001), the production process is subject to the specific word orders of the language. For languages with more rigid word orders, speakers might be more subject to syntactic constraints during online processing. On the other hand, languages with less rigid word orders allow speakers to focus more instead on the content of their utterances and how to deliver their ideas more efficiently and precisely. They might put long constituents first because they tend to be semantically more salient than short phrases. Hence when a long-before-short PP ordering is observed before the head verb, it could result from the longer PP encoding more semantic information.

On the other hand, in particular from a production-based point of view, if a speaker orders the two PP dependents based on DLM, that indicates the speaker has planned ahead what to say for both phrases, then figures out which one is shorter and puts it closer to the head verb (Chang, 2009). Though this is likely when both PPs are short, the scenario becomes less plausible when one or both PPs are relatively long. This advance planning seems even harder and against the incremental nature of langauge production when two PPs appear preverbally rather than postverbally, since the former requires

planning of not just the two PPs, but also what the head verb is and where it occurs. Previous production studies of head-final languages have demonstrated that there is not always advanced selection of verb during processing (Momma et al., 2016). Thus whether and under what context constituent orders abide more by “short closer" or “short first" deserves more thorough exploration.

3.4.3 The relationship between DLM and relative length difference

In his examination of the 394 instances in English, Hawkins (1999) demonstrated that the extent of DLM is modulated by the relative weight/length difference between the two PP dependents. As shown in Table 3.2 (adpated from Hawkins (1999)), as the length difference increases, the shorter PP is more likely to be closer to the head verb, and occurs first, which in turn shortens overall dependency length.

	$PP_2 > PP_1$ **by 1**	**by 2-4**	**by 5-6**	**by 7+**
$[\text{V } PP_1\ PP_2]$	60%	86%	94%	99%
$[\text{V } PP_2\ PP_1]$	40%	14%	6%	1%

Table 3.2: DLM predictions for the 394 instances from Hawkins (1999).

However, when the two PPs appear before the verb, there are possibly different observations. On one hand, given the pressure for DLM, as the length difference between the two PPs increases, there might be a stronger tendency for the shorter PP to appear closer. On the other hand, particularly from the perspective of language production, as discussed in Section 3.4.2, the preference for “short first" (MacDonald, 2013) would motivate the shorter PP to occur first, and farther away from the head verb. In this case, it is possible that the larger the length difference between the two PPs is, the stronger preference there would be for the shorter PP to appear first instead.

So could the relationship between the extent of DLM and the length difference between the two PPs offer a better explanation for the observed patterns here? To examine this, for each language, I first separated the dataset based on four scales of length difference in

Table 3.2. Then I calculated the proportion of shorter PPs appearing closer within each group. Estimates for confidence intervals were derived from bootstrapping with 1,000,000 iterations.[4]

As shown in Figure 3.5, within each subgraph, each bar from left to right represents the proportion of cases where the shorter PP is closer when the length of the two PPs differs by 1 token, 2-4 tokens, 5-6 tokens, or larger, respectively. For all the languages with head-initial PPs after the head verb, there appears to be a stronger effect for DLM as the length difference between the two PPs increases. The preference for the shorter PP to be closer to the head verb is the weakest when the length difference between the two PPs is just 1 token.

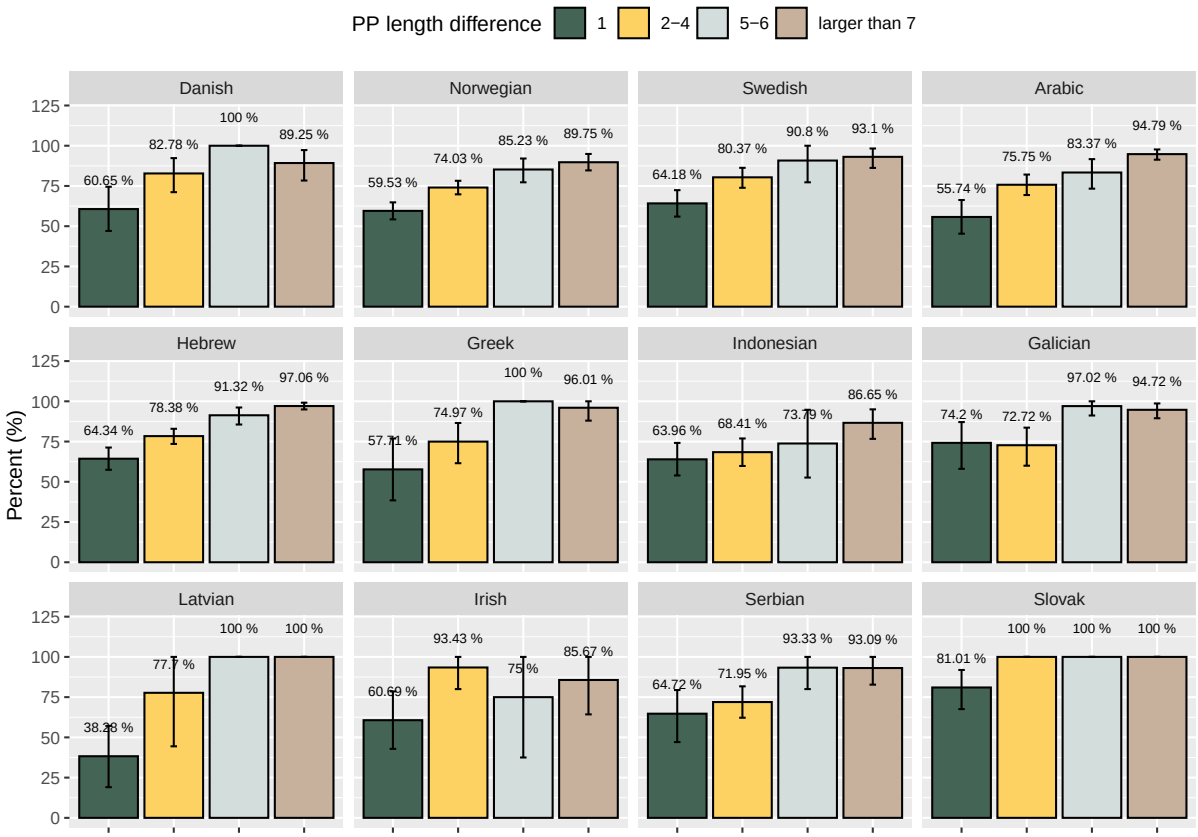


Figure 3.5: DLM given different relative weight in languages with head-initial PPs after the head verb. Error bars represent 95% confidence intervals.

[4]Same procedures were carried out for when choosing the content head within each PP and there were no observant differences in the results.

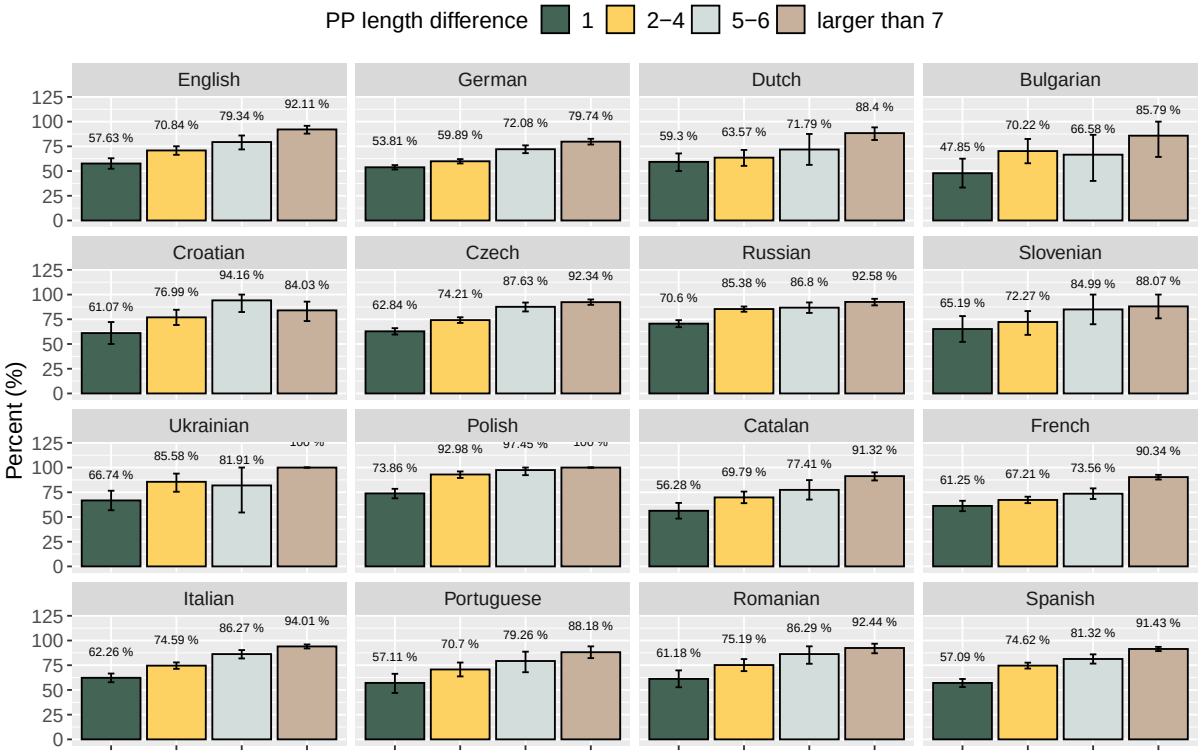


Figure 3.6: DLM given different relative weight when PPs appear after the verb, in mixed-type languages where PPs can also appear before the verb. Error bars represent 95% confidence intervals.

For the languages with head-initial PPs that can appear both after and before the head verb, when the two PPs are postverbal, as shown in Figure 3.6, I observe patterns that are comparable to Figure 3.5. The tendency for the shorter PP to occur closer is weak when the length of the two PPs differs by just 1 token; whereas there seems to be a more pronounced tendency for DLM as the length difference between the two PPs increases.

On the other hand, when the two PPs are before the head verb, regardless of its headedness (Figure 3.7, Figure 3.8, Figure 3.9), things are different. When two head-initial PPs appear before the head verb, in languages where PPs can also occur postverbally, as shown in Figure 3.7, there does not appear to be a stronger effect for DLM as the length difference between the two PPs increases in the Germanic and Romance languages, However, for the Slavic languages here, there seems to be a slightly more pronounced tendency for the shorter PP to appear closer when the length difference between the two

PPs is bigger.

In languages with only head-initial PPs before the head verb, as shown in Figure 3.8, again, there is no stronger preference for shorter dependencies with larger length difference between the two PPs. Similar patterns also hold for the three rigid OV languages here, in Figure 3.9.

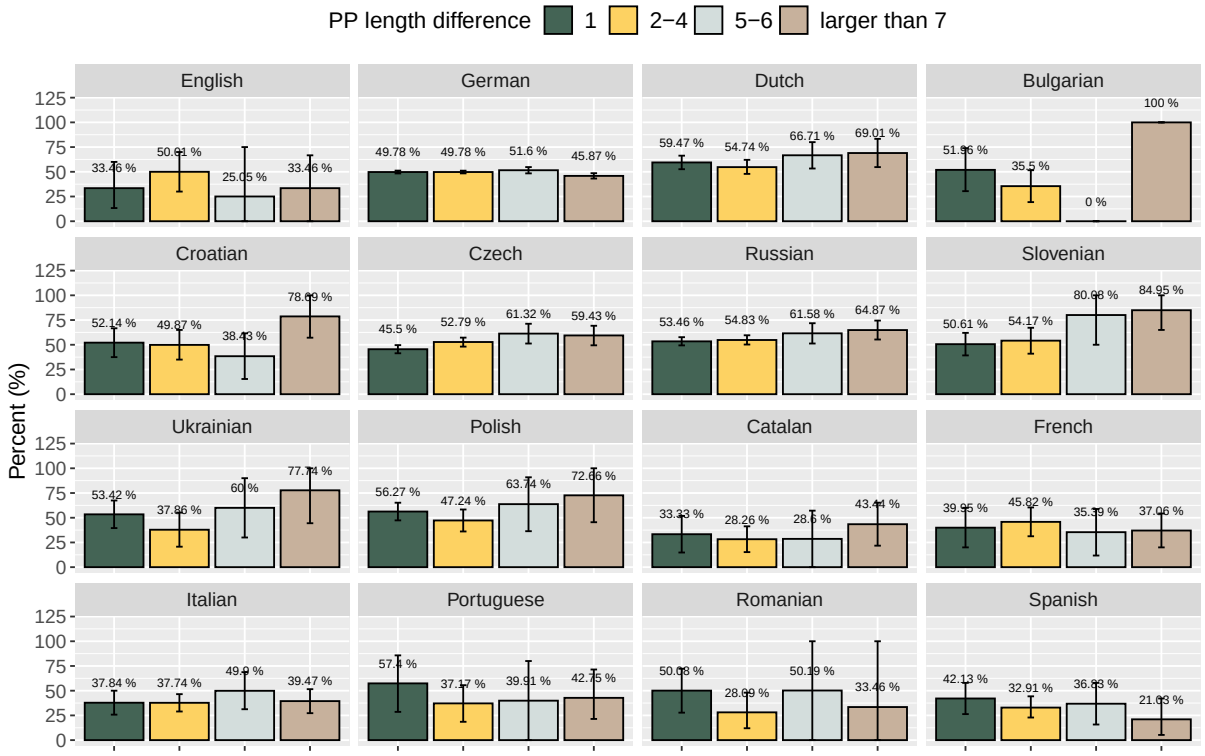


Figure 3.7: DLM given different relative weight when PPs appear before the verb, in mixed-type languages where PPs can also appear after the verb. Error bars represent 95% confidence intervals.

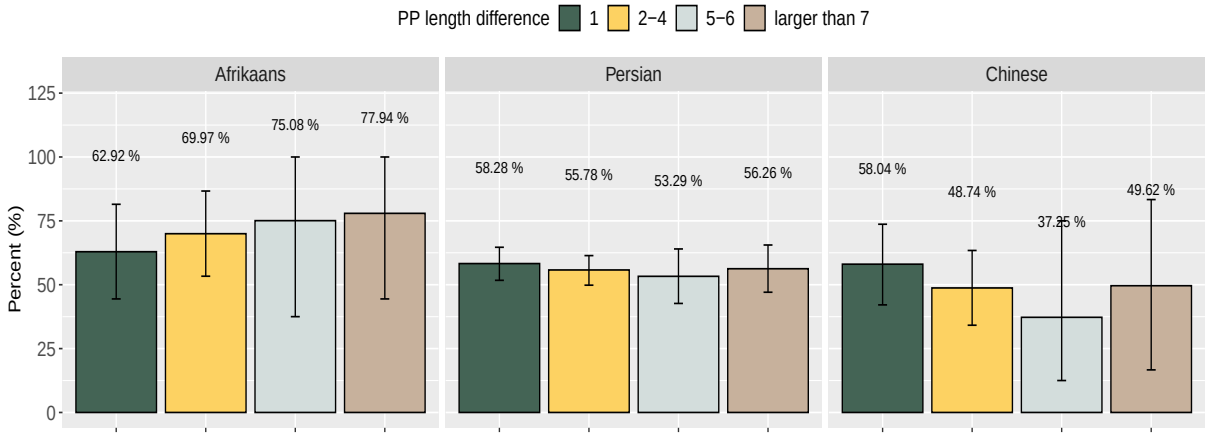


Figure 3.8: DLM given different relative weight in Afrikaans, Persian and Chinese. Error bars represent 95% confidence intervals.

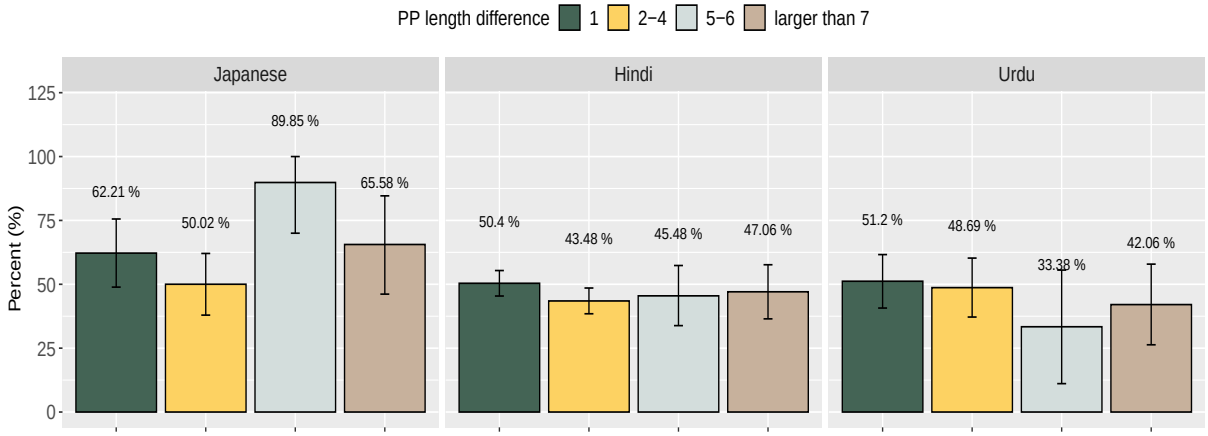


Figure 3.9: DLM given different relative weight in languages with head-final PPs before the head verb. Error bars represent 95% confidence intervals.

3.4.4 Other competing and cooperating factors

The results presented in both Section 3.3 and Section 3.4.3 together indicate there are other constraints, which are working conjointly with or pulling in different directions against dependency length to decide what the PP order will be. Accordingly, effects from other factors simultaneously lead to the different extents of DLM across languages.

Previous studies focusing on postverbal PP orders in English (Hawkins, 1999; Wiechmann and Lohmann, 2013) in particular have examined the roles of semantic closeness, discourse status and the traditional ordering rule of adverbial phrases, Manner Place

Time. Though they have only found mild or no effect for the latter two factors, both experiments have confirmed the significant predictive power of semantic closeness. These studies together have shown that in cases where dependency length has a weak or no effect, semantic closeness plays a stronger role by comparison. Thus it is possible there is a strong preference for the semantically closer PP to occur closer to the head verb across the languages here, regardless of or in a way that is less dependent on the dependency distance between the PP and the head verb.

In addition, the preverbal PP ordering patterns here corroborate evidence against DLM in adverb placement and preverbal modifying adjuncts in English from previous experiments (Gildea and Temperley, 2010; Temperley, 2007; Rasekh-Mahand et al., 2016). One of the reasons they have argued is that languages such as Chinese and Japanese tend to be more “topic-prominent", and they would prefer to put the topic phrase first before the head verb. Further studies are necessary to establish whether that is the case, which would in turn explain the weaker tendency for DLM in Chinese and Japanese found here.

Chapter 4

A Comparative Corpus Analysis of PP Orders in English and Chinese

4.1 Introduction

Chapter 3 found that crosslinguistically, the tendency for shorter dependencies is much less pronounced or does not exist at all in preverbal domains compared to that in postverbal domains. Additionally, this contrast is less dependent on the grammatical characteristics of specific language genera and it does not seem to be constrained by whether the PP is head-initial or head-final. For instance, the preference for DLM is much weaker or disappears both in languages with head-initial PPs before the head verb (e.g. Dutch, Russian) as well as in languages with head-final PPs before the head verb (e.g. Japanese, Urdu). By contrast, the effect for dependency length is the strongest in almost all languages with head-initial PPs after the verb except for Latvian. This indicates that there are other factors that operate together with dependency length to decide what the optimal orderings are, and these factors potentially play stronger roles than dependency length in determining the ordering preferences within preverbal contexts.

To examine what structural constraints cooperate and/or compete with dependency length and how effective they are in both preverbal and postverbal orders, this chapter presents a comparative analysis of constituent ordering preferences in English and Mandarin Chinese, two languages with distinct typological properties. Similar to Chapter 3, I used PP orderings as a test case, focusing on VP instances with exactly two PP

dependents appearing on the same side of the head verb, the ordering of which permits flexibility in at least some contexts. For both languages, I explored corpora with straightforward annotations that are larger than their treebanks in the Universal Dependencies project version 2.5 (UD) (Zeman et al., 2019) used in Chapter 3. I used the Penn Treebank (PTB) (Marcus et al., 1993) for English, and the Penn Chinese Treebank version 5 (CTB) (Xue et al., 2005) for Chinese. I probed to what extent dependency length, argumenthood status and the traditional adverbial ordering principle, Manner Place Time, compete and cooperate with each other in explaining the observed ordering preferences. I examined the effects of these three factors and their interactions across different genres of English, and between the two languages.

4.2 Experiments

4.2.1 Data

For data extraction, I followed the annotation system of both PTB and CTB. I searched for sentences with VP containing exactly two PPs on the same side of the same head verb, where the ordering of the PPs allows some flexibility under certain contexts. [1] All the PPs in the extracted instances have a P-NP structure. For English, all instances have a V-PP-PP structure; whereas for Chinese, all instances have a PP-PP-V structure (see Section 4.3.1).

Chapter 3 described three different possibilities for the comparatively complex adposition system in Chinese. As shown below, (1) is considered prepositional, (2) is considered postpositional, whereas there is debate regarding whether (3) should be considered prepositional or rather a circumposition (Cai, 2013; Zhang, 2013).

(1) 在　美国
in America
'in America'

(2) 书包　里
bag in
'in the bag'

[1] Manual inspection of a sample from each corpus suggested a large majority of sentences fit the criteria.

(3) 从 法庭 上
from court on
'from court'

In CTB, the post-nominal element in phrases like (3) is annotated as a localizer, and is combined with the preceding NP to form a localizer phrase. In this case, for example, 法庭 上 in (3) will be deemed a localizer phrase instead of an NP. Since every PP in my dataset has a P-NP structure, cases like (3) were not extracted and considered here.

Corpus	**Total**
WSJ	3596
Brown	3033
Switchboard	1187
CTB	250

Table 4.1: Total number of instances in which head verb has two PP dependents on the same side for each corpus.

4.2.2 Measures of each factor

4.2.2.1 Dependency length

The effect for dependency length was estimated in the same way as in Chapter 3. Confidence intervals were computed using bootstrapping (Efron, 1979) for 1,000,000 iterations for each language.

4.2.2.2 Argument status

To decide the argument status of a PP, I borrowed the coding scheme from Merlo and Ferrer (2006), which carefully distinguishes PP arguments and adjuncts given their annotated grammatical function and semantic tag from the treebanks, shown in Table 4.2. As described in their paper, the motivation for including untagged PPs as arguments was due to the fact that in the corpora, NPs (direct object & indirect object) and sentential constituents that are clearly arguments of the verb are left untagged (Marcus et al., 1994; Bies et al., 1995).

One thing to be noted here is that whether a constituent is an argument or an adjunct is not always definitive. In other words, it is not ideal to make a binary choice regarding

the argumenthood status of a constituent in relation to its syntactic head. Instead, the notion of argumenthood status should be considered as a gradient one. Therefore, rather than looking at each PP as strictly an argument or an adjunct, I interpreted the coding scheme by Merlo and Ferrer (2006) as a way to approximate how argument-like and adjunct-like each PP is relative to the head verb. To analyze the effect of argument status, I only examined VP instances with one argument-like PP and one adjunct-like PP (WSJ: $n = 1371$, Brown: $n = 1048$, Switchboard: $n = 470$, CTB: $n = 68$). I computed the proportion of cases in which the argument-like PP occurs closer and the proportion of cases in which the adjunct-like PP appears closer. Confidence intervals were estimated with bootstrapping for 1,000,000 iterations.

Argument-like PPs	
-CLR	dative object if dative shift not possible (e.g., donate); phrasal verbs; predication adjuncts
-EXT	marks adverbial phrases that describe the spatial extent of an activity
-PUT	locative complement of *put*
-DTV	dative object if dative shift possible (e.g., *give*)
-BNF	benefactive (dative object of *for*)
-PRD	non VP predicates
untagged PPs	
Adjunct-like PP	
-DIR	direction and trajectory
-LOC	location
-MNR	manner
-PRP	purpose and reason
-TMP	temporal phrases

Table 4.2: Grammatical functions and semantic tags of PP constituents in the treebanks.

4.2.2.3 Manner Place Time

In the treebanks, certain PPs have function tags that denote manner (PP-MNR), place (PP-LOC) or time (PP-TMP). I restricted analysis to sentences that have both PPs annotated with these function tags. For English, I calculated whether the order of the two PPs follows the ordering of Manner before Place before Time (MPT) (Quirk et al., 1985). As mentioned in Chapter 2, this rule is claimed to operate in the opposite direction in preverbal orderings in verb-final languages. Although Chinese is predominantly SVO, it is considered a mixed-type language in the typology literature. When a head verb has two PP dependents on the same side, as demonstrated in Chapter 3 (see also Section 4.3.1), they occur before the verb. Thus for Chinese, I computed whether the order of the two PPs follows the reverse ordering of Time before Place before Manner (TPM) instead.

4.2.3 Evaluating predictive power

I further compared the predictive power of dependency length and argument status in PP orderings with logistic regression modeling, which has been widely used to model structural preferences (Bresnan and Ford, 2010; Levy and Jaeger, 2007; Wasow et al., 2011). I did not include the rule of MPT in the model as the number of cases where it applies was quite small (see Section 4.3.5). I trained the regression models to predict the observed orders in the corpora over their structural variants (Rajkumar et al., 2016). In order to do this, I first randomly selected half of the original instances extracted from the corpora and left them the way they were. For the other half, I constructed their structural variants simply by switching the order of the two PPs. Hence in the dataset of each corpus, half of the sentences are the originals while the other half are the constructed variants.

The outcome variable was the ordering of the two PPs, a binary variable represented as *Order*. I coded *Order* as 1 for all original sentences, and 0 for all variants. Dependency length and argument status were included as the predictors in the model. For dependency length, I coded it as 1 when the shorter PP is closer to the head verb, -1 when the longer PP is closer, and 0 when the two PPs have the same length. For argument status, I coded it as 1 when the argument-like PP appears closer, -1 when the adjunct-like PP occurs

closer, and 0 when the argument status of the two PPs is the same. A summary of the coding scheme for the predictors is presented in Table 4.3. For each model, I evaluated its prediction accuracy with 10,000-fold cross-validations.

Factor	**1**	**-1**	**0**
dependency length	short PP closer	long PP closer	equal length
argument status	argument-like PP closer	adjunct-like PP closer	same argument status

Table 4.3: Coding for predictors in logistic regression models.

4.3 Results & Analysis

4.3.1 PP orderings in the two languages

For English, across the three corpora in PTB, when the head verb has two PP dependents on the same side, they appear as head-initial PPs after the verb, shown as below.

(4) Zoey presented [**on** something linguistic] [**to** her colleagues and friends] .

(5) Zoey presented [**to** her colleagues and friends] [**on** something linguistic] .

By contrast, in CTB, when a head verb has two PPs occurring on the same side, they appear as head-initial PPs before the head verb, shown in (6) and (7). Overall, the PP orderings for both languages here are the same as those observed in UD in Chapter 3.

(6) 他们 将 [在 生产 电子产品 方面] [和 中国] 进行 互利 合作

They will [in the aspects of electronic device production] [with China] conduct mutually beneficial collaboration.

(7) 他们 将 [和 中国] [在 生产 电子产品 方面] 进行 互利 合作

They will [with China] [in the aspects of electronic device production] conduct mutually beneficial collaboration.

'They will carry out collaborations that are mutually beneficial with China in the production of electronic devices.'

4.3.2 Effect of dependency length

As presented in Figure 4.1, the preference for DLM is substantial across genres in English. The number of sentences that have the shorter PP closer to the verb is 1.8 to 3.5 times larger than the number of sentences that have the longer PP closer to the verb. Although the numbers in the Switchboard corpus show a robust tendency for shorter dependencies, this tendency is comparatively weaker than that in both WSJ and Brown, indicating that the effect of dependency length is weaker in spoken data in contrast to that in written text. On the other hand, in roughly 20% of all sentences in the three corpora, dependency length makes no prediction, since the two PPs have the same number of tokens.

By contrast, preverbal PP orderings in Chinese only demonstrate a mild effect for dependency length. The number of cases when the shorter PP appears closer is not significantly much larger than that of instances where the longer PP is adjacent. This is slightly different from the observation in Chapter 3, where I found no preference for DLM

in Chinese at all. This is likely due to the different number of instances found in the two corpora (CTB: $n = 250$; Chinese UD treebank: $n = 111$). Overall, the preference for DLM in preverbal domains of Chinese is much weaker than that in postverbal domains of both written and spoken English.

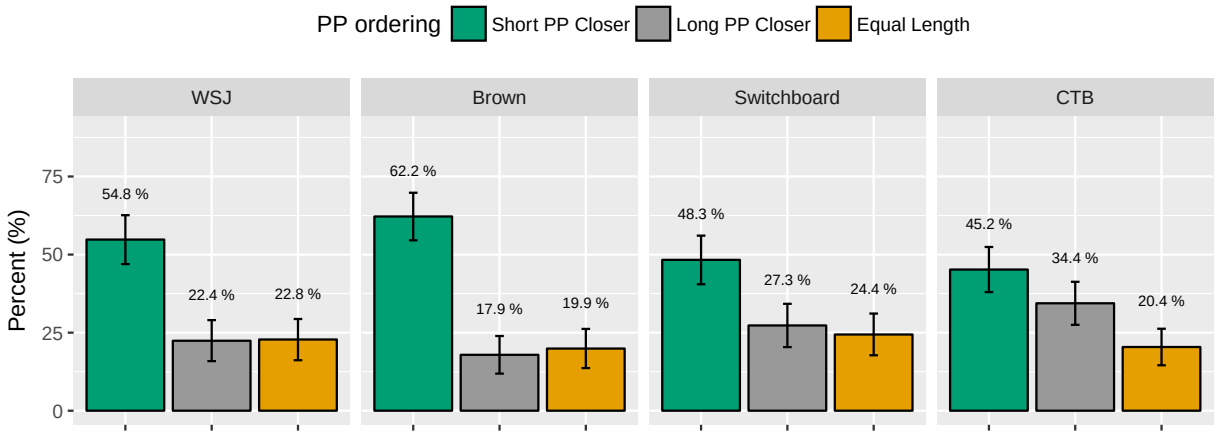


Figure 4.1: Effect of dependency length. Error bars represent 95% confidence intervals.

In order to better understand why DLM might be weaker in spoken genre of English than in written texts, I took a closer look at the lengths of the PPs in the datasets. I conjectured that there could be two possible reasons. First, compared to written texts, the average PP length for spoken English might be much shorter. Secondly, spoken data might have more cases where the length difference between the two PPs is relatively small. Both aspects can be expected to make it less necessary to put the shorter PP closer to the head verb in Switchboard. This is because from a psycholinguistic perspective in particular, in both of these cases the processing efficiency benefits derived from shorter dependency lengths are weaker, leading to an overall weaker preference for DLM.

To test these conjectures, I computed the average PP lengths as well as the number of cases where the lengths of the two PPs differ by only 1-2 words. However, as shown in Table 4.4, the average PP length in Switchboard is comparable to that in Brown, and only mildly shorter than that of WSJ (by 0.3 word). The proportion of cases where the two PPs have small length difference in Switchboard is similar to that in Brown, while slightly higher than that in WSJ (by 1.2%). This suggests that there are other

potential constraints possibly competing with dependency length and working in different directions. They play stronger roles in spoken than written domains in English and have overruled the impact of dependency length.

Corpus	Average PP length	% with small PP length difference
WSJ	5.4 (4.8, 6.0)	34.7 (28.0, 41.4)
Brown	4.7 (4.1, 5.3)	42.8 (35.9, 49.7)
Switchboard	4.0 (3.5, 4.5)	49.5 (42.6, 56.4)

Table 4.4: Comparisons of PP lengths with 95% confidence intervals.

4.3.3 Effect of argument status

Consistent across domains for English and for Chinese, there appears to be a strong preference for the more argument-like PP to be closer to the head verb. The number of instances where an argument-like PP is more adjacent is 1.5 to 2.7 times larger than when the adjunct-like PP occurs closer. Interestingly, argument status has a stronger effect in Switchboard and CTB than in WSJ and Brown; whereas as shown in Section 4.3.2, dependency length is less effective in Switchboard and CTB than in WSJ and Brown. This suggests that there is stronger competition between the two factors in both spoken English and Mandarin Chinese compared to that in written English.

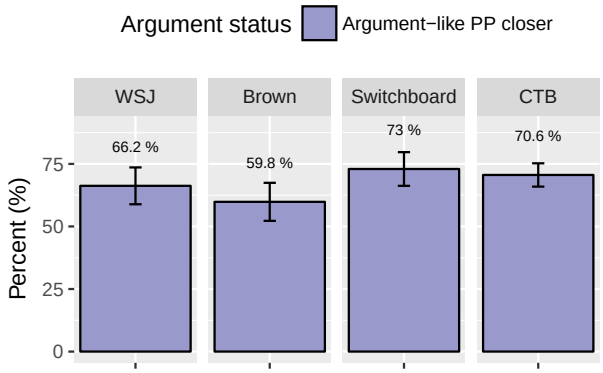


Figure 4.2: Effect of argument status of the PP. Error bars represent 95% confidence intervals.

Thereby I continued to ask how argument status interacts with dependency length in PP orderings. I estimated and compared the effect of argument status in sentences

where the shorter PP is closer versus when the longer PP is closer. [2] In particular, in cases where the shorter PPs are closer, it might matter less whether these shorter PPs are argument-like or not, since dependency length is already exerting a positive effect in favor of short-before-long ordering. By comparison, when the longer PPs are closer, it is possible that most of the longer PPs are arguments of the verb in order to minimize the lexical domains, as proposed in MiD by Hawkins (2004) (see Section 2.2.1.1).

Though the results shown in Figure 4.3 do not align exactly with my initial expectations, I do observe some interesting patterns. Across the genres for English, when the longer PPs appear closer, most of these longer PPs are more argument-like. More specifically, in WSJ, when the longer PPs are closer, the number of cases where the more argument-like PP prefers proximity to the head verb is much higher than that when the shorter PPs are closer. In addition, the numbers also suggest that in WSJ, there is a strong preference for DLM even when the shorter PP is more adjunct-like. This means that in WSJ, dependency length and argument status might have comparable predictive power in deciding what the PP ordering will be. On the other hand, in both Brown and Switchboard, the argument-like PP is more adjacent to the verb regardless of whether it is the shorter or the longer PP. The consistently pronounced effect for argument status here indicates that it possibly bears a stronger role than dependency length (Wiechmann and Lohmann, 2013; Marblestone, 2007) in these two corpora. When the two factors are pulling in different directions, the PP order might abide more by predictions of argument status than of dependency length.

4.3.4 Cooperation and competition between dependency length and argument status

To further compare the cooperation and competition between dependency length and argument status, I quantified the predictive power of the two factors with logistic regression models. I examined cases where at least one of the two constraints is effective. This means that instances where the two PPs have the same length as well as the same argumenthood

[2] Chinese was not included due to the small size of its data, which does not allow for meaningful statistical comparison of these two different cases.

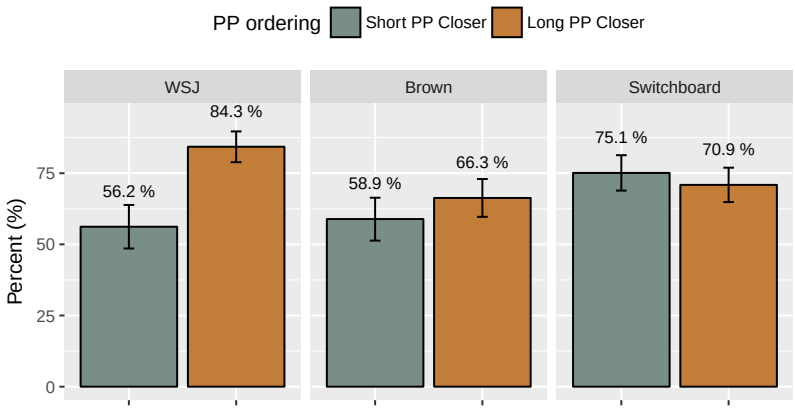


Figure 4.3: Effect of argument status when short vs. long PP is closer. Error bars represent 95% confidence intervals.

status were not included. Results from Figure 4.4 demonstrate that dependency length and argument status cooperate as well as compete with each other to different extents in the treebanks. The relative strengths of the two factors vary across domains in English and across the two languages.

The most strongly preferred order is when the PP that is both shorter and argument-like is adjacent to the head verb. On the other hand, competition between the two factors arises when they pull in the opposite directions (i.e. when the shorter PP behaves more like an adjunct or when the longer PP is more argument-like). The comparable predictive power for the two constraints in WSJ speaks to what I suggested earlier (see Section 4.3.3). This also indicates that there is strong competition when dependency length and argument status are working against each other. In Brown, dependency length appears to be more predictive than argument status, showing that the shorter PP is still more likely to be closer even when it is not an argument. In other words, the PP orderings in Brown will align more with predictions by DLM.

In both Switchboard and CTB, argument status has a more pronounced role than dependency length. This contrast indicates that in these two corpora, the orders tend to put the argument-like PP adjacent to the head verb, even if it is the longer PP between the two PPs. In other words, the weaker extent of DLM found in both the spoken genre of English as well as Chinese possibly results from a stronger effect for argument status

in the two domains.

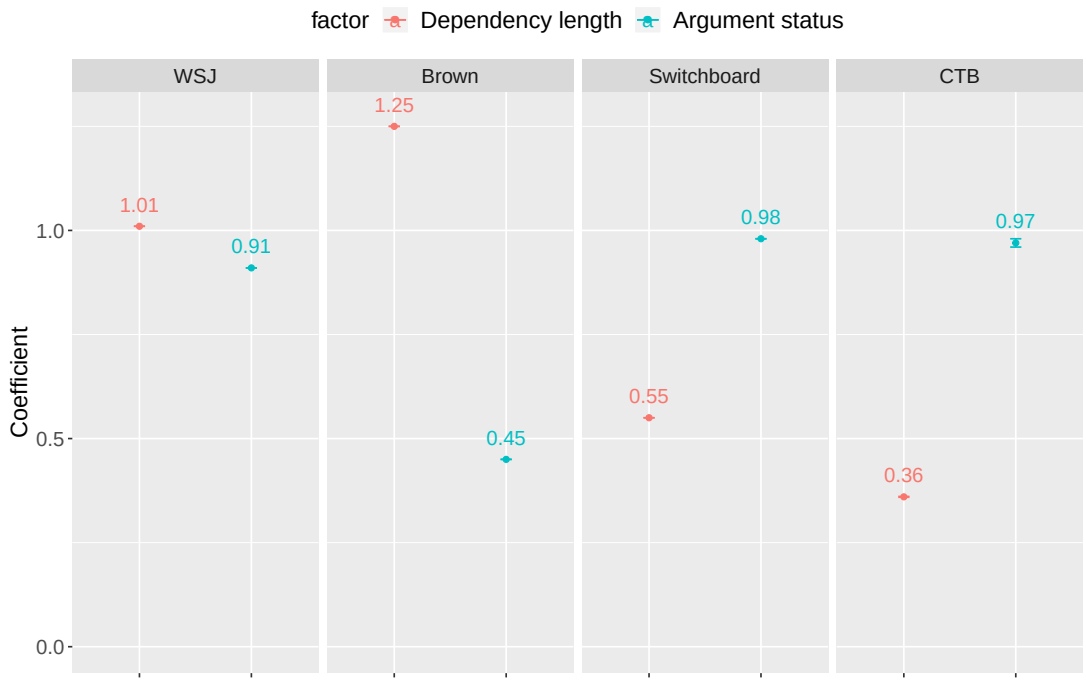


Figure 4.4: Coefficient estimates for dependency length and argument status in each corpus. Dots show point estimates with error bars representing 95% confidence intervals.

4.3.5 Effect of Manner Place Time

Though there were not enough cases where the Manner Place Time principle applies to Chinese, I found a significant role for this ordering principle in English across genres. In the current dataset, Manner Place Time applies to about 6% of all instances. Within this set, it correctly accounts for the order of 89.3% of sentences in WSJ, 100% in Brown, and 100% in Switchboard. However, because it applies so infrequently, its overall impact is much smaller than that of dependency length and argument status. Similarly, previous studies from Hawkins (1999) and Wiechmann and Lohmann (2013) showed no or only weak evidence for the Manner Place Time principle (though with smaller data samples) operating in addition to dependency length and argumenthood in PP orders.

4.4 Discussion

In this chapter I have analyzed the effects of dependency length, argument status and Manner Place Time in PP orderings from both English and Chinese. Consistent with previous studies, dependency length serves as a strong predictor for postverbal PP orderings across domains in English, though its effect is much weaker in the spoken genre in comparison to written text. Nevertheless, it only exerts a mild effect at best in preverbal PP orders in Chinese.

On the other hand, argumenthood status has a pronounced effect in both languages. It appears to have comparable or even stronger predictive power than dependency length in predicting ordering preferences. This observation is in line with corpus studies from Wiechmann and Lohmann (2013) and Hawkins (1999) as well as with psycholinguistic experiments from Marblestone (2007), all looking at PP orders in English. In particular, the role of argument status is much more pronounced in postveral PP orderings in spoken English and in preveral PP orderings in Chinese, which explains the weaker preference for DLM observed in both contexts.

Corpus	**Accuracy (%)**
WSJ	67.3 ± 0.03
Brown	73.4 ± 0.03
Switchboard	66.4 ± 0.04
CTB	63.3 ± 0.10

Table 4.5: Model prediction accuracy with dependency length and argument status.

A natural follow-up study would be to ask whether more argument-like or semantically closer constituents tend to be adjacent to their syntactic heads on a larger crosslinguistic scale. As presented in this chapter, it is possible that there is a significant effect for argumenthood status across languages, which would offer an explanation for the mixed evidence for DLM in preverbal vs. postverbal domains found in Chapter 3. In addition, as effective as dependency length and argument status are in this chapter, it is clear that around 30% of the data in English and around 40% of the data in Chinese remain

unexplained based on the model prediction accuracy from Table 5.3. Other constraints and their interactions with dependency length and argument status await to be discovered. I explore these two directions further in Chapter 5.

Chapter 5

Quantifying Lexical and Structural Constraints in Crosslinguistic Constituent Ordering Preferences

5.1 Introduction

While findings from both Chapter 3 and various previous studies (Gulordava et al., 2015; Faghiri and Samvelian, 2020; Yamashita and Chang, 2001) have shown a preference for DLM within syntactic alternations across languages, the different extents of DLM in different languages as well as between preverbal and postverbal domains typologically call for explorations of other cooperating and competing biases that are effective and operate simultaneously with dependency length crosslinguistically. Chapter 4 gave a focused analysis comparing constituent ordering preferences in English and Mandarin Chinese, using double PP orderings as the test case. In English, when a head verb has two PP dependents on the same side, they appear postverbally; whereas in Mandarin Chinese, the two PPs occur preverbally. I have presented evidence that an explanation for the contrast between English and Chinese is that there seems to be a strong tendency for the more argument-like phrase to appear closer to the syntactic head, regardless of whether that will yield overall shorter dependencies.

Nevertheless, the results from the aforementioned two chapters leave a few important questions unanswered. First, the effect of argument status in Chapter 4 was measured

based on gold standard manual annotations, which are not available in a large crosslinguistic context. It is unclear whether argument status or semantic closeness is an effective typological determinant of ordering preferences. Secondly, what other structural motivations are at play in predicting syntactic choices across languages and how to measure them computationally need more exploration. Thirdly, it remains to be seen whether dependency length still has strong predictive power when the roles of other factors are controlled for.

This chapter takes up these questions. I investigated the roles of four additional lexical and structural factors in predicting constituent ordering preferences across languages with different grammatical characteristics. The four factors are semantic closeness, lexical frequency, contextual predictability and word co-occurrence information. Using PP order typology described in Chapter 3 as a test case, I demonstrated how to quantify these constraints with computational methods. I further compared the individual predictive power of these four constraints to that of dependency length.

Except for dependency length and semantic closeness, previous research has rarely addressed the direct roles of lexical frequency, contextual predictability and word co-occurrence information in constituent orderings in a crosslinguistic context. Nevertheless, the effects of these three factors in language processing are well motivated. If the reason for why language users prefer DLM lies in the fact that shorter dependencies in general ease processing, then I would expect lexical frequency, contextual predictability and word co-occurrence information, which have been shown to facilitate processing efficiency, to affect syntactic orderings as well.

5.2 Motivation

5.2.1 Lexical frequency

The correlation between frequency and structural complexity can be traced back to the markedness hierarchies proposed by Greenberg (1966). For instance, in languages with rich morphology, the markedness hierarchy of case (Nom>Acc>Dat>Other) reflects the frequency ranking of the different cases. In other words, as the formal marking goes

from *Nom* to *Other*, the frequency of occurrence for each case declines. The markedness hierarchy of number (Sing>Plur>Dual>Trial/Paucal) shows the same correspondence to frequency (e.g. in English the singular form *dog* occurs much more frequently than the plural form *dogs*, and is morphologically less complex than the plural form). As suggested by Keenan and Comrie (1977) in their Accessibility Hierarchy, the underlying cause for such patterns is attributed to ease of processing which declines for each position down the hierarchy. The processing load of different hierarchy positions is shaped by their complexity and frequency of occurrence. More frequent categories are associated with greater processing ease, accessibility and predictability, whereas less frequent items are harder to access; they require more effort for activation and processing, and more explicit coding is needed down the hierarchies.

Recent corpus-based studies and psycholinguistic experiments have provided more empirical evidence for frequency-based accounts for language processing (Jurafsky, 2003). It has been shown that language users exploit frequency information of linguistic structure at different levels for both language comprehension and production (Arnon and Snider, 2010; Arnon, 2015). The processing preferences for different structures are shaped by how frequently they occur in a language (Thornton et al., 1998). For example, Bybee and Hopper (2001) showed that subjects identify or recognize more frequent words much faster. Jurafsky (1996) demonstrated that the frequency of different verb subcategorization frames influences syntactic ambiguity resolution. Previous work has found that the processing of multi-word expressions is shaped by the overall frequency of the expressions (Arnon and Snider, 2010; Morgan and Levy, 2016a). More frequent linguistic items tend to be phonetically reduced more easily in order to ease production effort (Jurafsky et al., 2001).

Previous studies have presented evidence for the effect of lexical frequency in constituent orderings, though they have largely focused on constructions on the word level and have mainly examined English. For instance, it has been shown extensively that the more frequent word tends to appear first in binomial expressions in English (Morgan and Levy, 2015; Benor and Levy, 2006; Fenk-Oczlon, 1989; Gustafsson, 1976). Other studies

have demonstrated similar patterns for adjective orderings in English (Trotzke and Wittenberg, 2019; Wulff and Gries, 2015; Wulff, 2003). Here I asked whether lexical frequency predicts PP ordering typology, i.e., whether the PP with more frequent individual words tends to appear first. I approximated the lexical frequency of each PP as the product of the frequency of each word within the PP (see Section 5.4.2.2). From the perspective of production ease in particular, the more frequent PP is more likely to occur first since it is easier to retrieve or access on average (Jaeger and Tily, 2011; Christianson and Ferreira, 2005; Kempen and Harbusch, 2004).

5.2.2 Contextual predictability

Though more frequent linguistic elements in general are relatively easier to process, the effect of lexical frequency appears to be modulated by the preceding context (Ehrlich and Rayner, 1981). Given a fixed preceding context, the linguistic element that is more predictable is more likely to facilitate processing (Ferreira and Lowder, 2016; Jurafsky, 1996; Kutas et al., 2011; Kuperberg and Jaeger, 2016). One of the most salient models that advocate for the role of context is the expectation-based processing model (Levy, 2008), a variant of the broader experience-based approach to processing (MacDonald et al., 1994), which has mainly been developed to account for comprehension phenomena.

To show the contrasting effect of lexical frequency and contextual predictability, let us consider the following examples:

(1) I had some *bread*.

(2) I had some *fries*.

(3) I had some burgers and *bread*.

(4) I had some burgers and *fries*.

In the neutral context of (1) and (2), the word *bread* which has higher frequency on average should be processed faster compared to the word *fries*, which is less frequent. Nevertheless, given the preceding context *I had some burgers and*, I would expect that (4) is easier to process in comparison to (3).

Previous work looking at the role of contextual predictability in syntactic alternations has mainly focused on syntactic reduction in spoken English, demonstrating that the more predictable a linguistic unit is, the more likely it is to be reduced in production. For instance, existing findings have shown that whether a relative clause (RC) in English starts with a relativizer or not (e.g. *that is the sweetest thing* ***that*** *I have ever seen* vs. *that is the sweetest thing* ∅ *I have ever seen*) depends on both the lexical predictability of *that* as well as the structural predictability of an RC given preceding context (Wasow et al., 2011; Jaeger and Levy, 2007; Jaeger, 2010). Since lexical frequency discards context completely, I also investigated the role of contextual predictability in PP ordering typology, and whether the more predictable PP appears first given previous context. I approximated the contextual predictability of each PP as the product of the predictability of each individual word within the PP given preceding context (see Section 5.4.2.3).

5.2.3 Word co-occurrence information

Recent empirical evidence has demonstrated that words that are more likely to co-occur with each other in contexts are words that depend on each other statistically, and that these head-dependent pairs are more likely to be placed closer to each other for ease of processing (Futrell and Levy, 2017; Futrell et al., 2017). This has been formulated as the general principle *information locality* (Futrell et al., 2020b). In the context of word order preferences, information locality predicts that when a syntactic head has multiple dependents with flexible orderings, the dependents that have higher (pointwise) mutual information with the head tend to appear closer. This principle has so far been applied in explaining adjective orderings in English (Futrell, 2019; Hahn et al., 2018). Here I explored whether word co-occurrence information plays a role in PP ordering, and whether the PP which is more likely to co-occur with the head verb would be placed closer. I approximated the co-occurrence information between each PP and the verb as the pointwise mutual information between the nominal head of each PP and the verb (see Section 5.4.2.4).

5.3 Predictions of the four factors

As presented in Chapter 3, the typological patterns of PP orderings depend on the word order features of different languages. Given the predictions of the four factors described above, it is clear that all of dependency length, semantic closeness and word co-occurrence information always pull in the same direction across languages, regardless of the PP ordering patterns of the language. In other words, whether the language has two PPs both appearing after or before the head word, the PP that is closer to the head verb in the sequence of the two PPs will be preferred to be shorter, semantically closer and more likely to co-occur with the head verb. This applies whether the PP is head-initial, as in Hebrew, or head-final, as in Japanese.

Nevertheless, the effect of lexical frequency and that of contextual predictability might be modulated by the specific PP ordering structure. If shorter PPs tend to be more frequent and more predictable given preceding context, for languages with two PPs after the head verb like Indonesian, a preference for shorter dependencies will place the shorter PP closer to the verb, and this means that presumably the more frequent or the more predictable PP will occur first. In this case, lexical frequency or contextual predictability will be cooperating with dependency length and pulling orders in the same direction as well.

On the other hand, for languages with two PPs before the head verb, if the shorter PP is in fact more frequent or more predictable, the preference for DLM would violate the expectation that the PP of higher frequency or higher contextual predictability should occur first. In this case, lexical frequency and contextual predictability are no longer in agreement with dependency length regarding PP ordering. Instead, they would be pulling in different directions from the effect of dependency length, which possibly leads to a weaker preference for DLM. Furthermore, this should also apply whether the PP is prepositional or postpositional.

5.4 Experiments

This section describes the experiments carried out to quantify the effects of semantic closeness, lexical frequency, contextual predictability and word co-occurrence information. I asked whether there is a crosslinguistic tendency for the more frequent or the more predictable PP to appear first, for the PP that is semantically closer or more likely to co-occur with the head verb to be placed closer. I further investigated the predictive power of these four factors along with dependency length with regression modeling.

5.4.1 Data

Estimates of lexical frequency, contextual predictability and word co-occurrence information in particular (see Section 5.4.2) need large training data for each language. I used UD-style treebanks from the raw data of the CoNLL 2017 Shared Task on multilingual parsing (Ginter et al., 2017). These treebanks contain texts from both Common Crawl and Wikipedia and are automatically parsed with UDPipe (Straka and Straková, 2017). The initial 34 languages in Chapter 3 were mostly Indo-European (29), with the exceptions of Afro-Asiatic (2), Austronesian (1), Japanese (1) and Sino-Tibetan (1). I selected all languages that are not Indo-European and picked a representative sample of 16 Indo-European languages due to data availability. This resulted in a dataset with 21 languages in total. For Hindi, Urdu and Afrkaans, I only measured the effects of dependency length and semantic closeness due to limited amount of training data for the three languages. For the remaining 18 languages, I calculated the effects of all five factors.

5.4.2 Measures of each factor

5.4.2.1 Semantic closeness

Examinations of semantic closeness or argumenthood status on a crosslinguistic scale are rare, despite the recognition of its significant role in ordering preferences. The precise meaning of semantic closeness as well as the ideal way to compute it have not been fully established. In this chapter, I approximated semantic closeness between each PP and the head verb by measuring how semantically similar the head verb and the nominal head of each PP are. To do this, I used cosine similarity and word embeddings, which have been

Languages	**Language family and genus**
Languages with head-initial PPs after the head verb	
Arabic	Afro-Asiatic, Semitic
Greek	Indo-European, Greek
Hebrew	Afro-Asiatic, Semitic
Indonesian	Austronesian, Malayo-Sumbawan
Slovak	Indo-European, (West) Slavic
Languages with head-initial PPs after or before the head verb	
English	Indo-European, Germanic
Dutch	Indo-European, Germanic
Croatian	Indo-European, (South) Slavic
Slovenian	Indo-European, (South) Slavic
Polish	Indo-European, (West) Slavic
Czech	Indo-European, (West) Slavic
Russian	Indo-European, (East) Slavic
Ukrainian	Indo-European, (East) Slavic
Italian	Indo-European, Romance
Spanish	Indo-European, Romance
Afrikaans	Indo-European, Germanic
Persian	Indo-European, Iranian
Chinese	Sino-Tibetan
Languages with head-final PPs before the head verb	
Japanese	Japanese
Hindi	Indo-European, Indic
Urdu	Indo-European, Indic

Table 5.1: Information for selected languages.

widely applied in computational linguistics. I considered the PP with a nominal head that is semantically more similar to its head verb to be the PP that is semantically closer to the verb, or more argument-like.

Word embeddings are high-dimensional distributed vector representations for words or phrases (Schütze, 1993; Heinzerling and Strube, 2018; Levy and Goldberg, 2014a,b; Garg et al., 2018; Hamilton et al., 2016; Mikolov et al., 2013a,b; Pennington et al., 2014). Each word is mapped onto a vector space of numerical values, one numerical value for each dimension. To approximate the semantic similarity between two words, I calculated the similarity between the two vectors, or the two embeddings that represent those two words using cosine similarity, a metric that measures the cosine of the angle between the two vectors. Let $\vec{v_1}$ and $\vec{v_2}$ represent the vectors for two different words. Their cosine similarity is computed as:

$$cosine(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}||\vec{v_2}|} \tag{5.1}$$

Imagine each of the three words *love*, *cheese*, *kale* has three-dimensional embeddings as follows:

$$\begin{aligned} love&: \left[\ 5,\ 8,\ 6\right] \\ cheese&: \left[\ 6,\ 3,\ 5\right] \\ kale&: \left[\ 7,\ 4,\ 2\right] \end{aligned}$$

Then the cosine similarity between *love* and *cheese* as well as *love* and *kale* is computed respectively as:

$$cos(love, cheese) = \frac{5 \times 6 + 8 \times 3 + 6 \times 5}{\sqrt{5^2 + 8^2 + 6^2}\sqrt{6^2 + 3^2 + 5^2}} = 0.90 \tag{5.2}$$

$$cos(love, kale) = \frac{5 \times 7 + 8 \times 4 + 6 \times 2}{\sqrt{5^2 + 8^2 + 6^2}\sqrt{7^2 + 4^2 + 2^2}} = 0.85 \tag{5.3}$$

Based on the results above, *love* is semantically more similar to *cheese* than *kale*.

I used word embeddings from Multilingual BERT (Devlin et al., 2019), a single neural language model pre-trained from monolingual corpora of 104 languages. For every VP instance, I initially calculated the semantic similarity between the head verb and the nominal head within each PP for every VP instance using both contextual and non-contextual word embeddings derived from BERT. [1] Contextual embeddings for a word are represen-

[1] One alternative estimate would be to measure the semantic closeness between the verb and the whole PP, not just the nominal head, although this is sensitive to the length of the PP. Therefore I did not experiment with this measure here.

tations of the word generated from the model taking both its preceding and following sentential context into consideration, while non-contextual embeddings are representations of just the word by itself generated from the model. I chose contexutal embeddings eventually.

5.4.2.2 Lexical frequency

To approximate the lexical frequency of a PP, I first estimated the probability of each token within a PP from the training data for each language. The probability of the PP was then calculated as the product of the probability for each token within that PP. Since the product of the probabilities is clearly sensitive to the length of the PP, to separate the contributions of PP length and lexical frequency, I looked instead at the perplexity of each PP, which is the inverse probability of the PP, normalized by the number of tokens contained within the PP. For a PP with a sequence of tokens $W = w_1 \dots w_n$, its perplexity is calculated as:

$$Perplexity(W = w_1 \dots w_n) = P(w_1 w_2 ... w_n)^{-\frac{1}{n}} \tag{5.4}$$

Thus based on Eq. 5.4, for instance, the perplexity of the PP *in the morning* is computed as follows:

$$Perplexity(in\ the\ morning) = (P(in) * P(the) * P(morning))^{-\frac{1}{3}} \tag{5.5}$$

According to the prediction that more frequent items are produced first, the PP with lower perplexity should be the first one in the sequence of the two PPs.

5.4.2.3 Contextual predictability

I measured the contextual predictability of each PP with estimates from neural language models (LM). Specifically, I used long-short term memory models (LSTMs) (Hochreiter and Schmidhuber, 1997), which generate word-by-word predictions. The architecture of the LM was the same for every language. I extracted 50 million tokens for each language from their training data, shuffled them by sentences and split them into training, development and test sets (8:1:1). I employed standard 2-layer word-level LSTM with tied weights for the input layer and the output layer. Every model had 650 units for the hidden

layers, a batch size of 64, and a dropout rate of 0.2. I trained each model for 40 iterations, with early stopping based on performance on the development set. The implementation was in Pytorch. [2]

Concerning contextual predictability, neural models with more efficient architecture than the simple LSTMs trained here might yield better predictions. The reason I did not use multilingual BERT to also generate word-by-word prediction is due to its original training objective, which is that the target word is predicted by both preceding and following context. Since I am only interested in the predictability of each PP given preceding context, the motivation for which is driven by the literature on human language processing, I opted to train models that would only generate word-by-word prediction from the preceding context. Additionally, a recent study by van Schijndel et al. (2019) compared the performance of simple LSTMs versus large-scale neural transformer models such as BERT and GPT on the tasks of syntactic agreement. The results showed that transformer-based models do not always have better prediction accuracy in certain constructions even when they are trained on much larger datasets.

Similar to the calculation of lexical frequency, the contextual predictability of a PP was the product of the conditional probability of every token within that PP given preceding sentential context, which was generated from the softmax layer of the LSTM. Again to separate the effects of PP length and predictability, for a PP with a sequence of tokens $W = w_1 \dots w_N$, I measured its perplexity given previous context as Eq. 5.6.

$$Contextual\ Perplexity(W = w_1 \dots w_n) = \left[\Pi_1^n P(w_n | w_{1 \dots n-1})\right]^{-\frac{1}{n}} \tag{5.6}$$

Given Eq. 5.6, the contextual predictability of each of the two PPs in the sentence *Zoey sang* $\left[_{PP_1}\ \textit{with the band}\ \right]$ $\left[_{PP_2}\ \textit{on Tuesday}\ \right]$ is calculated respectively as follows:

$$Contextual\ Perplexity(with\ the\ band) = \left[P(with|Zoey\ sang) * P(the|Zoey\ sang\ with) * P(band|Zoey\ sang\ with\ the)\right]^{-\frac{1}{3}} \tag{5.7}$$

[2] `https://github.com/pytorch/examples/tree/master/word_language_model.`

$$Contextual\,Perplexity(on\,Tuesday) = \left[P(on|Zoey\,sang)*P(Tuesday|Zoey\,sang\,on)\right]^{-\frac{1}{2}} \tag{5.8}$$

5.4.2.4 Word co-occurrence information

I used pointwise mutual information (PMI) to measure word co-occurrence information. PMI computes how often two words w_1 and w_2 co-occur, compared with what would be expected if independence between the occurrence of each of the two words is assumed (Jurafsky and Martin, 2019).

$$pmi(w_1, w_2) = log\frac{p(w_1, w_2)}{p(w_1)p(w_2)} \tag{5.9}$$

I derived PMI from the training data of each language. For each VP instance, I measured the PMI of the head verb and the nominal head within each PP. I regarded the PP with a nominal head that has higher PMI with the head verb to be the PP that is more likely to co-occur with the verb.

5.4.3 Evaluating predictive power

I computed and compared the predictive power of each individual factor in PP orderings with Bayesian mixed-effects logistic regression modeling. [3] I initially experimented with two model architectures: one that includes the data for all the languages and one that was fit for each individual language. The two model types had the same fixed effects, which were the five factors along with the pronominality of the head noun in each PP. The random effects in the first type were the head verb in each instance, language family and the ordering structure of each language [4], with random intercepts and random slopes for each fixed effect for the ordering structure of every language. The random effect in the second was the head verb in each instance. To better handle issues of missing data induced by fitting one model that includes data for all the languages, I eventually fit one

[3]Results were replicated and validated using linear logistic mixed models with the same model architecture.

[4]For example, for VP instances where the two PPs occur before the verb in German, the ordering structure was coded as *German preverbal*.

model with the same architecture for every language. The coding scheme for each fixed effect is presented in Table 5.2.

```
Order ~ dependency length + semantic closeness + lexical frequency +
   ↪ predictability + co-occurrence information + pronominality
            + (1|Verb)
```

Factor	**1**	**-1**	**0**
dependency length	short PP closer	long PP closer	equal length
semantic closeness	semantically similar PP closer	semantically similar PP farther	equal semantic closeness
lexical frequency	more frequent PP first	less frequent PP first	equal frequency
contextual predictability	more predictable PP first	less predictable PP first	equal predictability
word co-occurrence information	more likely to co-occur PP closer	more likely to co-occur PP farther	equal co-occurrence

Table 5.2: Coding for fixed effects in logistic regression models.

For each language, I randomly selected half of the original sentences and kept them the way they were; for the remaining half I constructed their structural variants by exchanging the order of the two PPs. In the dataset of each language, half of the sentences are the originals while the other half are the constructed variants. The outcome binary variable was the ordering of the two PPs, *Order*. I coded *Order* as 1 for all original instances, and 0 for all variants.

I trained the model to predict the original order. I used a weakly informative prior for

each parameter in the model (Hahn et al., 2018; Ghosh et al., 2018), which is helpful to set reasonable upper and lower bounds for values within the posterior distributions (Levshina, 2018). The prior followed a Student's t distribution centered around 0 (mean $\mu = 0$), with $\nu = 3$ degrees of freedom and scale $\sigma = 10$. I employed the default Markov Chain Monte Carlo (MCMC) sampling methods implemented in STAN using the R package **brms** (Bürkner et al., 2017). I ran 50 chains with 2000 iterations in each chain, with the first 200 iterations as warmup (burn-in) samples. Confidence intervals for each model parameter were derived from their respective posterior distributions.

5.5 Results

5.5.1 Languages with head-initial PPs after head verb

Among the 21 languages, for Arabic, Greek, Hebrew, Indonesian and Slovak, when a head verb has two PP dependents appearing on the same side, they appear as head-initial PPs after the head verb. Based on results from Chapter 3, all of these five languages have demonstrated a strong preference for DLM. As shown in Figure 5.1, when the effects of the other four factors are controlled for, there still appears to be a robust effect for dependency length.

By comparison, the predictive power of dependency length is the strongest among the five constraints. With that being said, semantic closeness appears to have a positive effect in the ordering preferences. There seems to be a tendency for the semantically closer or more argument-like PP to be adjacent to the head verb. On the other hand, based on Figure 5.1, none of the other factors appears to have a consistently significant effect across the five languages.

5.5.2 Languages with head-initial PPs after or before head verb

Within the dataset, ten languages have mixed PP orderings. That is, for these languages, when a head verb has two PP dependents on the same side, they can appear as head-initial PPs both after and before the head verb.

As shown in Figure 5.2, when the PPs appear after the head verb, dependency length is an effective predictor across the ten languages. There is a strong preference for the shorter

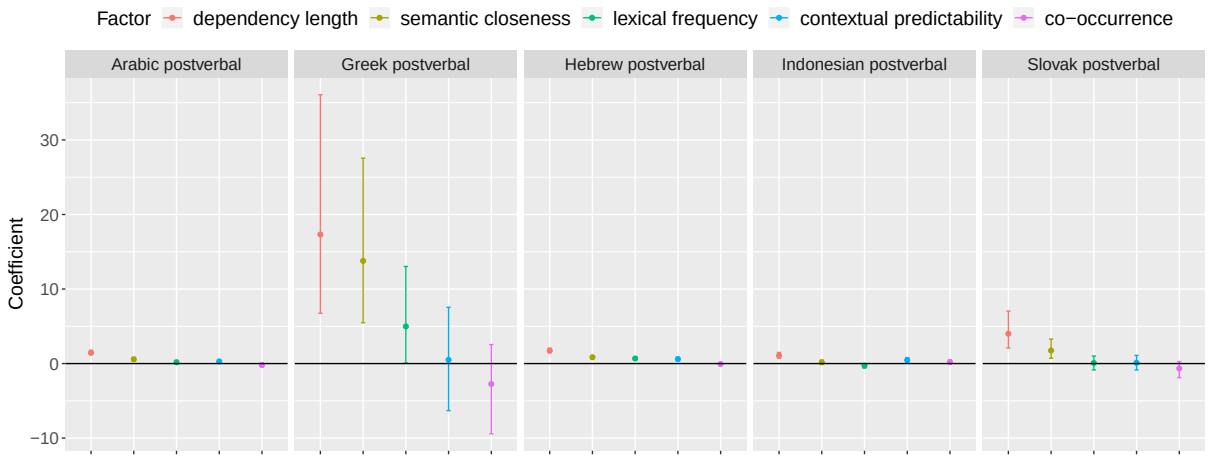


Figure 5.1: Coefficient estimates for the five factors in languages with head-initial PPs after the head verb. Error bars represent 95% confidence intervals.

PP to appear closer to the head verb. Semantic closeness also seems to have a significant effect in all the languages, though to a lesser extent than dependency length except for English. The stronger role of semantic closeness in English corresponds to previous findings from Hawkins (1999) and Wiechmann and Lohmann (2013) on PP orders, both of which showed that when the effects of other interacting principles are controlled for, lexical dependency is the most predictive. By contrast, the roles of the other three constraints are much weaker and are far from being relatively consistent. For most of the languages in general, none of lexical frequency, contextual predictability and co-occurrence information appears to be predictive at all.

On the other hand, when the PPs occur before the head verb, as presented in Figure 5.3, the effect of dependency length still disappears even when the roles of other constraints are controlled for. This lends more support to the original findings from Chapter 3 where just by looking at dependency length, most of these languages did not show a strong pattern for or against DLM. Semantic closeness appears to have a weak effect in five languages here: Dutch, Slovenian, Polish, Czech and Russian (β and 95% confidence intervals for semantic closeness in these languages are significantly higher than 0). Although in general, none of the five factors appears to play a consistent role in PP orderings across the ten languages.

Besides the ten languages seen above, three other languages, Afrikaans, Persian and

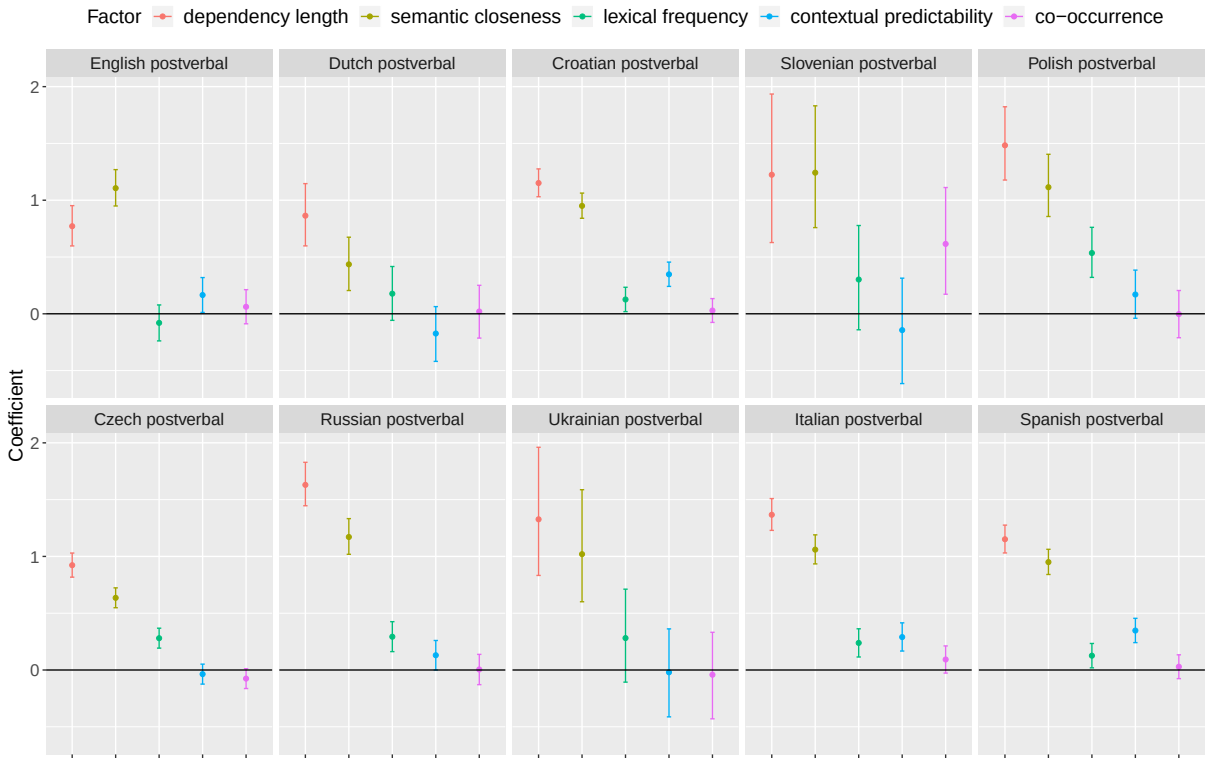


Figure 5.2: Coefficient estimates for the five factors in languages with head-initial PPs after the head verb, where the PPs can also appear before the head verb. Error bars represent 95% confidence intervals.

Chinese only have head-initial PPs before the head verb when the verb has two PP dependents on the same side. For these three languages, dependency length has a robust effect in both Afrikaans and Chinese, while it has no predictive power in Persian. On the other hand, semantic closeness appears to have a comparable or stronger effect than dependency length in the three language (for Persian, $\beta = 0.44$, 95% confidence interval is (0.29, 0.59)). This means that the PP orders in the three languages prefer the PP that is more argument-like to be closer to the verb. By comparison, the roles of the other three constraints are not observant. Overall, the pattern for dependency length in Afrikaans corresponds to the preference for DLM found in Afrikaans in Chapter 3; the more significant role of semantic closeness than dependency length in Chinese aligns well with the results from Chapter 4 using data from the Penn Chinese Treebank version 5 (Xue et al., 2005).

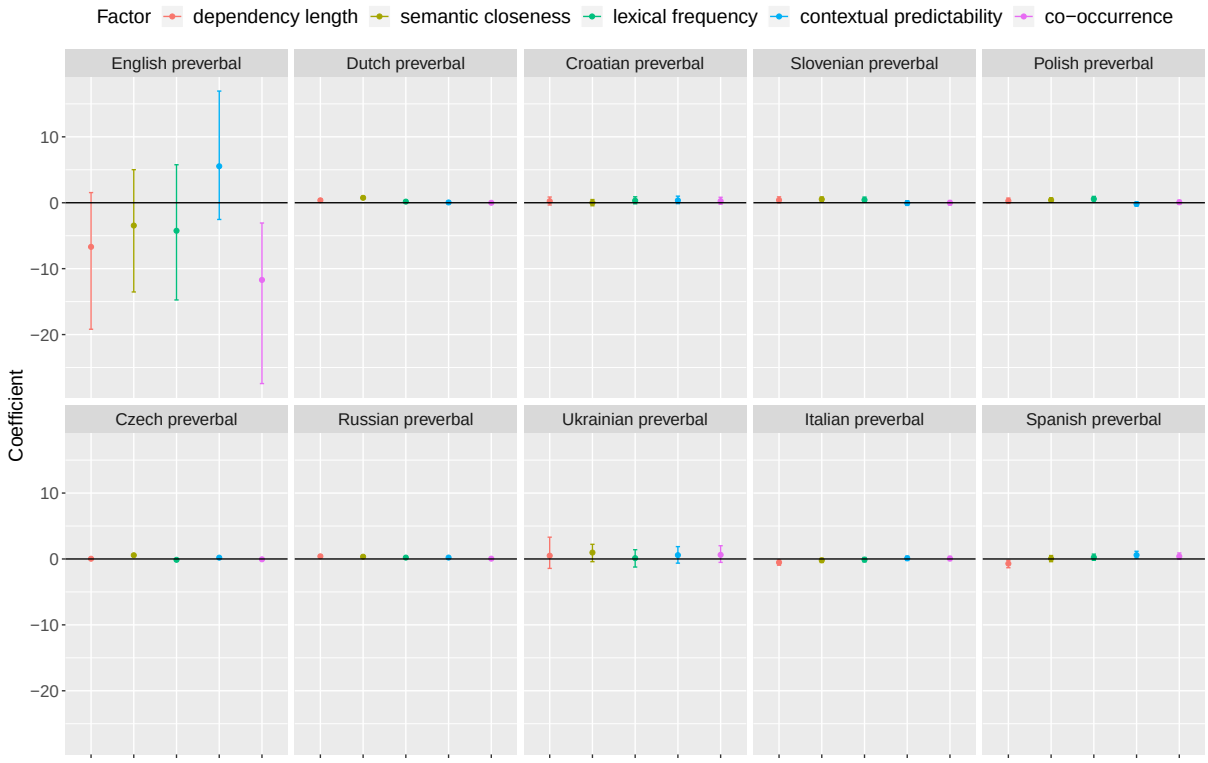


Figure 5.3: Coefficient estimates for the five factors in languages with head-initial PPs before the head verb, where the PPs can also appear after the head verb. Error bars represent 95% confidence intervals.

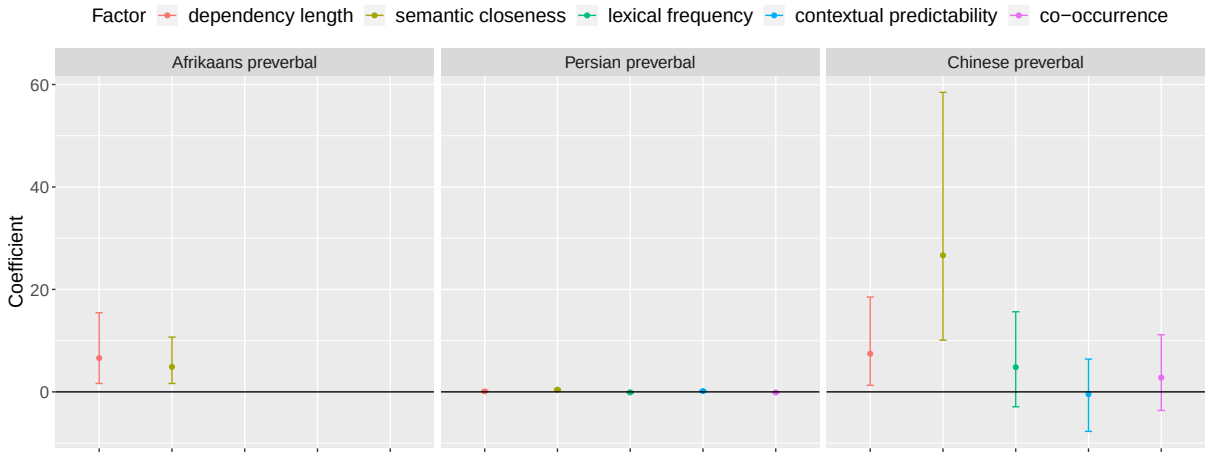


Figure 5.4: Coefficient estimates for the five factors in languages with head-initial PPs before the head verb. Error bars represent 95% confidence intervals.

5.5.3 Languages with head-final PPs before head verb

The last three languages to be analyzed in the dataset are the three rigid OV languages, Japanese, Hindi and Urdu. As shown in Figure 5.5, after controlling the effects of the other four factors, dependency length has no predictive power at all. This holds even in Japanese, whereas Chapter 3 as well as previous studies on Japanese (Yamashita and Chang, 2001) and Korean (Choi, 2007) have shown a preference for DLM in these rigid OV languages. There are two possible explanations for this discrepancy. First, this chapter is investigating the predictive power of dependency length when the effect for other factors is controlled for, while previous work and Chapter 3 were interested in dependency length as a sole predictor in ordering preferences. Secondly, the syntactic structure in question in this chapter is double PP construction, while previous work has focused on major constituent orders including the transitive and ditransitive constructions. As discussed in Chapter 3, the order of the PPs is possibly less subject to grammatical constraints than the order of major constituents such as subject, direct and indirect object, and this allows us to examine the role of dependency length more directly. With that being said, the different results presented here indicate that though there is a crosslinguistic tendency for DLM, its extent varies across different construction types.

On the other hand, semantic closeness has a significant role across the three languages and is the most predictive among the five constraints. This indicates that in the three rigid OV languages here, argumenthood status is the most pronounced factor in deciding the order of the two PPs.

5.6 Discussion

In this chapter, I demonstrated potential ways to quantify the effects of four lexical and structural constraints, namely semantic closeness, lexical frequency, contextual predictability and word co-occurrence information. I further compared their predictive power to that of dependency length with Bayesian mixed-effects modeling.

Overall in comparison, dependency length is the most predictive among the five constraints investigated in this chapter, indicating a crosslinguistic tendency for the shorter

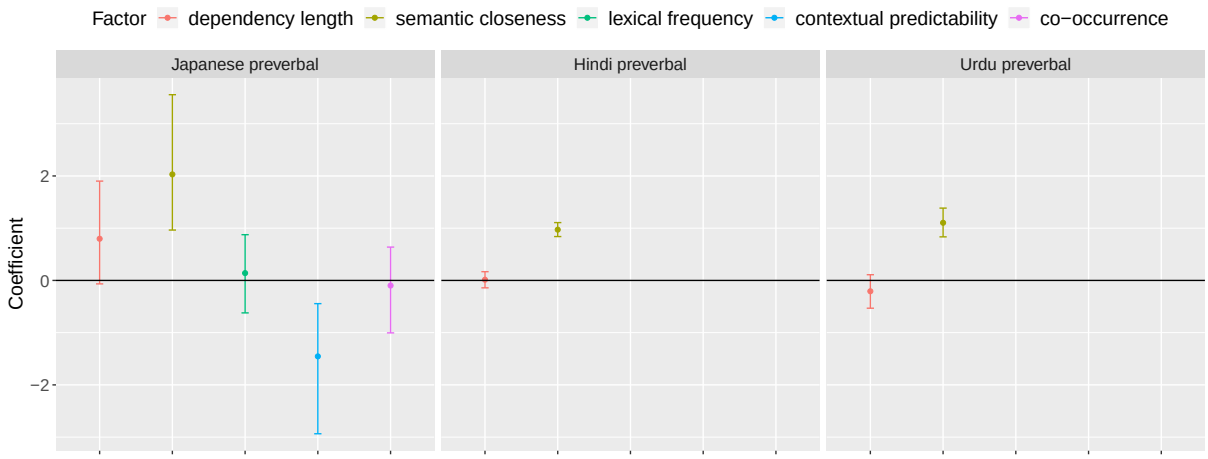


Figure 5.5: Coefficient estimates for the five factors in languages with head-final PPs before the head verb. Error bars represent 95% confidence intervals.

constituent to appear closer to their syntactic heads on the phrasal level. Though the predictive power of dependency length is much weaker or does not exist in preverbal domains in contrast to postverbal domains. On the other hand, semantic closeness also seems to have a significant and more consistent effect in most of the languages here, showing a typological pattern for more argument-like phrases to be adjacent to their syntactic heads. In cases where dependency length is not effective or only plays a weak role, semantic closeness appears to have a stronger effect, especially in preverbal domains than postverbal domains. This is a contrast that has not been quantitatively documented in previous literature.

There is simply no consistent preference regarding where the PP that is more frequent, more predictable, or more likely to co-occur with the head verb appears across these languages. The PP orderings are more subject to constraints such as dependency length and semantic closeness instead. In a post-hoc analysis, for each language, I have evaluated the prediction accuracy of models including different sets of factors with logistic regression modeling. Results (Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9, Figure 5.10) have affirmed that none of lexical frequency, contextual predictability and co-occurrence information boosts prediction accuracy significantly.

It may well be that there are other alternative ways of measuring lexical frequency, contextual predictability and co-occurrence information. For instance, instead of comput-

ing contextual predictability of each PP as the product of the predictability of every word within the PP, it can be estimated instead as just the product of the predictability of the adposition and the head noun within the PP. Alternatively, it may be that the methods employed here are reliable and that these three factors do not in fact play a significant role in at least the PP orderings that I have investigated, whereas dependency length and semantic closeness do. Previous work on double PP orderings (Hawkins, 1999; Wiechmann and Lohmann, 2013) has shown that when the other constraints are controlled for, givenness (more *given* item is more predictable given discourse context) has no or only a weak effect in determining the relative order of the two PPs. I leave further investigation of frequency-based predictions for syntactic orderings such as these to future studies. In addition, in Chapter 7 I return to discuss the role of frequency using a different syntactic structure: the dative construction in English.

Model	**Factors**
$Model1$	dependency length
$Model2$	dependency length + semantic closeness
$Model3$	dependency length + semantic closeness + lexical frequency
$Model4$	dependency length + semantic closeness + lexical frequency + contextual predictability
$Model5$	dependency length + semantic closeness + lexical frequency + contextual predictability + word co-occurrence information
$Model6$	dependency length + semantic closeness + lexical frequency + contextual predictability + word co-occurrence information + pronominality
$Model7$	dependency length + semantic closeness + contextual predictability + word co-occurrence information + pronominality + head verb as a random effect

Table 5.3: Different model architectures.

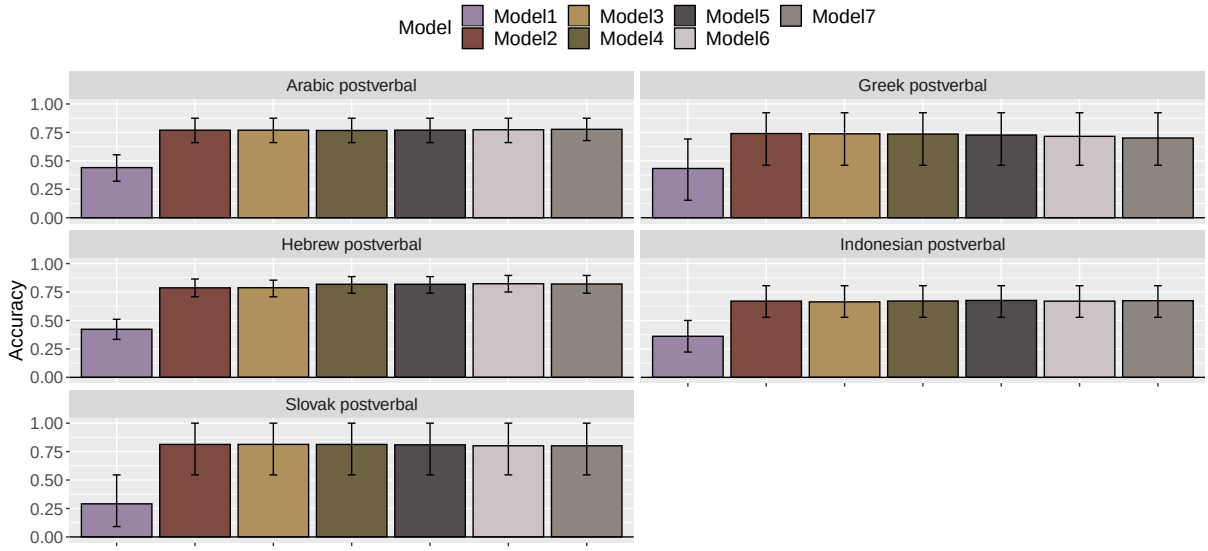


Figure 5.6: Model prediction accuracy in languages with head-initial PPs after the head verb. Error bars represent 95% confidence intervals.

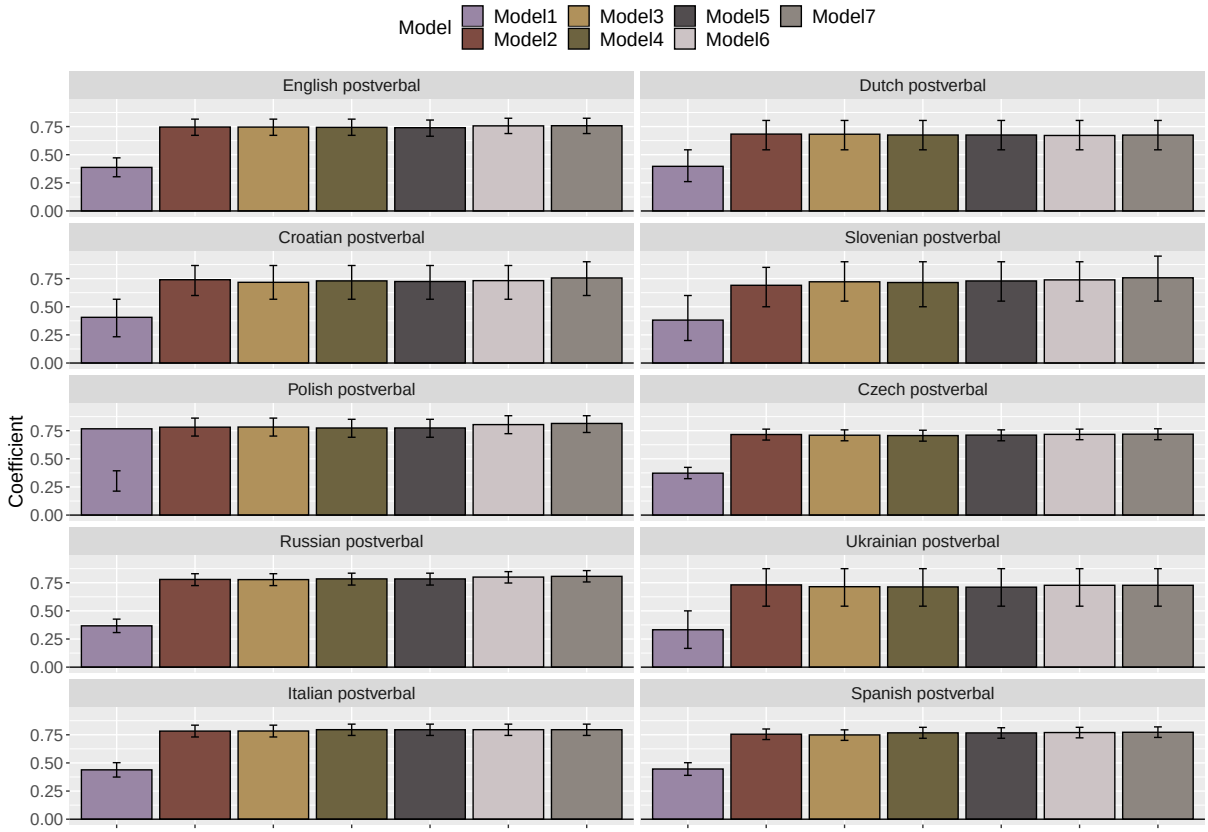


Figure 5.7: Model prediction accuracy in languages with head-initial PPs after the head verb, where the PPs can also appear before the head verb. Error bars represent 95% confidence intervals.

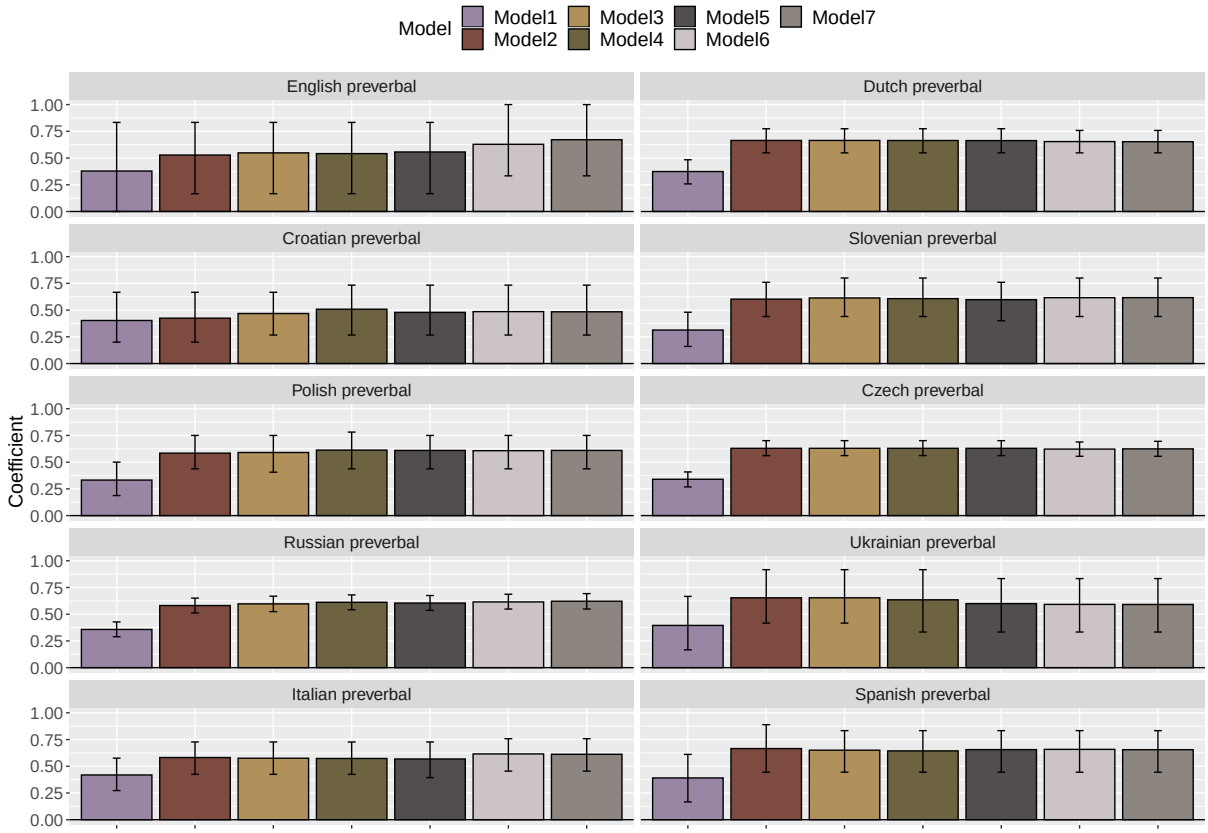


Figure 5.8: Model prediction accuracy in languages with head-initial PPs before the head verb, where the PPs can also appear after the head verb. Error bars represent 95% confidence intervals.

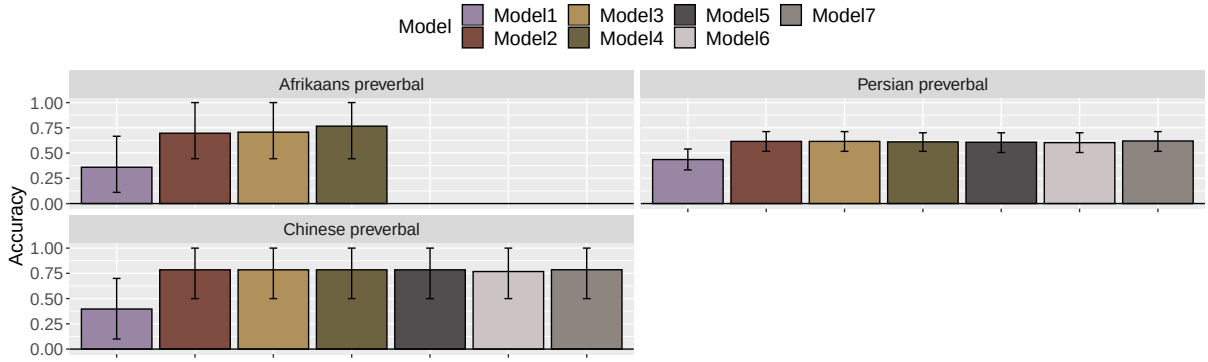


Figure 5.9: Model prediction accuracy in languages with head-initial PPs before the head verb. Error bars represent 95% confidence intervals.

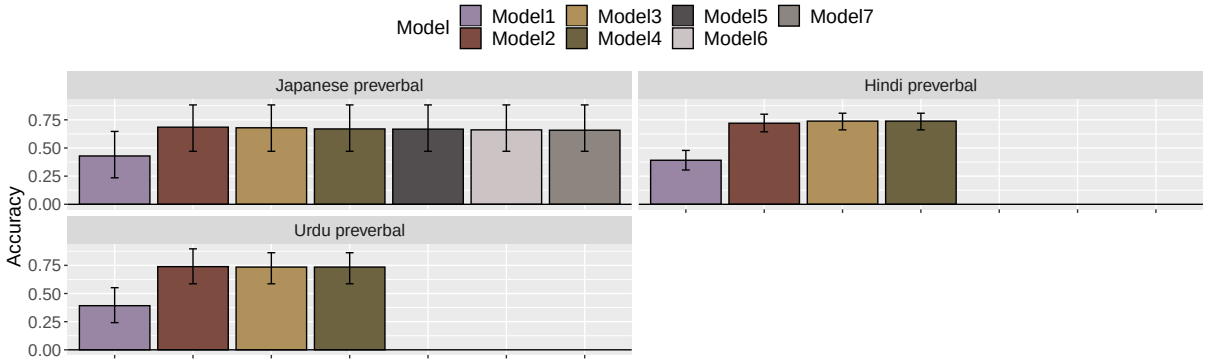


Figure 5.10: Model prediction accuracy in languages with head-final PPs before the head verb. Error bars represent 95% confidence intervals.

Chapter 6

The Crosslinguistic Relationship between DLM and Ordering Flexibility

6.1 Introduction

Overall, the results from Chapter 3 to Chapter 5 have demonstrated that there are relatively stronger preferences for DLM in postverbal constituent orderings, whereas there is no or only a weak tendency for shorter dependencies in preverbal orders. While I have tried to account for this pattern discrepancy by examining a number of different structural constraints that have been well-motivated theoretically, some recent studies have proposed another potential explanation from a different perspective: word order variability (Gildea and Temperley, 2010; Futrell et al., 2015a). These studies have suggested that the fact that certain languages have longer dependencies is due to their higher degree of word order freedom. When there is more flexibility in the orders of the constituents within a syntactic construction, the ordering preferences can possibly abide less by DLM and be more subject to other structural motivations. In this context, it could be the case that preverbal domains have more word order freedom than postverbal domains, leading to overall longer dependencies. On the other hand, it is also possible that these structures might take advantage of this greater ordering freedom by positioning shorter constituents closer to their syntactic heads in accordance with DLM.

In the literature on word orders and language typology, the notion of flexibility has not been given the attention it deserves. For decades, the categorization of the flexibility

profile of a full language usually relies on the relative order of main constituents such as subject, object and/or indirect object (Siewierska, 1998). The notion of flexibility has mainly been described as a categorical feature when it comes to characterizing syntactic properties of particular languages (Hale, 1982) (see also Namboodiripad (2019); Levshina (2019)). For instance, English is considered comparatively "fixed" (Polinsky, 2012); German is deemed much more flexible than English and is cast as "non-rigid OV" (Hawkins, 1986); Japanese is treated as "rigid OV", with presumably more ordering flexibility than languages like English or Swedish but less freedom than languages like German (Dryer, 2007).

Ordering flexibility should, however, be considered as a gradient word order parameter rather than a categorical one, and the characterizations of whether a language is "free" or "rigid" should take into consideration the particular syntactic constructions in question and the specific grammatical and discourse-level roles that these constructions take upon. For instance, when serving as a nominal modifier, adjectives in English, a predominantly SVO language, regularly appear before their head nouns.

(1) **reasonable** argument

(2) **good** point

Also in English, as demonstrated in the following examples, the adverbial phrase *yesterday* can appear relatively more freely, whether it occurs before or after its syntactic head *raised*, without yielding a significant difference in the interpretation of each sentence.

(3) **Yesterday** I think I raised a good point in Chapter 6.

(4) I think **yesterday** I raised a good point in Chapter 6.

(5) I think I raised a good point **yesterday** in Chapter 6.

(6) I think I raised a good point in Chapter 6 **yesterday**.

On the other hand, in Mandarin Chinese, a mixed-type language, when a sentence has the adverbial phrase 昨天, which also means *yesterday*, things are different. As shown in the following examples, while 昨天 has some flexibility when occurring before its syntactic head, 读了, it is rarely placed after the head verb.

(7) 昨天 我 读了 一本 书
zuotian wo dule yiben shu
yesterday I read-Past one-Classifier book

(8) 我 昨天 读了 一本 书
wo zuotian dule yiben shu
I yesterday read-Past one-Classifier book

(9) *我 读了 一本 书 昨天
wo dule yiben shu zuotian
I read-Past one-Classifier book yesterday

'Yesterday I read a book.'

In general, empirical investigations of word order flexibility as a way to classify and characterize languages have been lacking (Gulordava and Merlo, 2015a). The relationship between ordering variability and other word order properties needs more exploration. Recently there have been some attempts to bridge this gap. Futrell et al. (2015b) have quantified the degree of ordering freedom across 34 languages with multilingual corpora, focusing on the flexibility of a language's full word order profile. Their results corresponded well with previous grammatical descriptions of the word orders of the languages examined. Using diachronic corpora from Latin and Ancient Greek, Gulordava and Merlo (2015a) proposed different measures for the flexibility of modifiers (numerals and adjectives) and head noun within NPs. Their results have confirmed that the word orders of both languages have become more fixed over time. Levshina (2019) focused more on the ordering flexibility of particular syntactic dependencies and their grammatical functions both within and across languages, also using multilingual data. She demonstrated that empirical measures of ordering flexibility can be incorporated into the task of language classifications better than more traditional approaches based on language type. Taking a behavioral experimental approach, Namboodiripad (2019) compared the ordering variability of major constituents in English and Malayalam. She measured the ordering flexibility of the six permutations of subject (S), object (O) and verb (V) with acceptability judgement tasks. The results showed that the acceptability rating differences between canonical (SVO in English and SOV in Malayalam) and non-canonical orders are much

more significant in English than in Malayalam. Additionally, the experiments revealed word order properties of the two languages that have not been documented or considered theoretically before. For instance, based on the acceptability ratings, OVS in English is not considered as that “bad". And when the verb does not occur in the sentential-final position in Malayam, SVO and OVS orders are comparably acceptable.

This chapter makes a contribution towards this line of research. Specifically, I explored the *gradient* relationship between word order variability and DLM. I asked: do syntactic constructions with more ordering flexibility have a weaker preference for DLM across languages?

6.2 Experiments

6.2.1 The test case

To address the question, it is necessary to select syntactic structures that are crosslinguistically comparable and also exhibit different degrees of word order freedom across languages. To do this, first consider the following examples in English.

(10) Kobe **praised** $[_{NP}$ his oldest daughter $]$ $[_{PP}$ from the stands $]$.

(11) Kobe **praised** $[_{PP}$ from the stands $]$ $[_{NP}$ his oldest daughter $]$.

In both (10) and (11), the head verb *praised* has a direct object NP dependent and a PP oblique dependent, both occurring after the head verb. The two sentences are truth-conditionally equivalent, in the sense that they are semantically invariant and switching the order of the NP and the PP does not change the grammaticality of the sentence. I consider the direct object NP to be more *argument-like* (Merlo and Ferrer, 2006) in comparison to the PP. By the Principle of Argument Closer (see Section 2.2.2) (Culicover and Jackendoff, 2005; Staub and Clifton, 2006; Stallings et al., 1998), the structure of (10) will be more preferred to that of (11).

Nevertheless, in his influential book, Hawkins (2014) proposed the Principle of Argument Precedence as one of several interacting principles determining the relative ordering and flexibility of clause-level constituents. This principle states that more argument-like linguistic elements tend to occur before other dependents of the same syntactic head. The

motivation behind this principle lies in processing efficiency, and it is reasoned that more argument-like phrases are more accessible given sentential context. As noted in Hawkins (2014), this principle provides a valid explanation for why languages with a dominant word order of (S)OVX (X representing any oblique phrases) are more prevalent, while languages with a dominant order of (S)XVO are rare. The direct object is considered to be more argument-like compared to oblique phrases when they are governed by the same syntactic head, and an (S)OVX ordering follows the Principle of Argument Precedence.

Now let us reconsider the English examples given above. Both the direct object NP and the PP occur postverbally. Therefore when the NP is closer to the head verb, it also precedes the PP. In this case the Principle of Argument Closer and the Principle of Argument Precedence will be cooperating with each other and make the same prediction regarding the order of the two dependents. Accordingly, in a language with postverbal NP and PP dependents headed by the same verb, there should be mostly V-NP-PP orderings, i.e. a strong tendency for the NP to be adjacent to the verb as well as to be before the PP. In other words, the order of the NP and the PP dependent will accordingly be more fixed in this case.

However, OV languages such as Japanese have the mirror ordering pattern to that in English. When a head verb has both a direct object NP and a PP dependent appearing on the same side, they tend to appear preverbally. As illustrated below, both sentences are also truth-conditionally equivalent and moving the NP in front of the PP does not change the grammaticality nor the meaning of the sentence. Again I consider the direct object NP to be more argument-like than the PP. If the more argument-like phrase prefers proximity to the head verb, the structure of (12) will be preferred to that of (13).

(12) [$_{PP}$ 朝食 で] [$_{NP}$ パン を] 食べる
chōshoku de pan o taberu
(I) breakfast for bread -ACC eat

(13) [$_{NP}$ パン を] [$_{PP}$ 朝食 で] 食べる
pan o chōshoku de taberu
(I) bread -ACC breakfast for eat
'I eat bread for breakfast.'

Note that in contrast to English, when the NP occurs closer to the verb in Japanese, the PP precedes the NP instead. While the Principle of Argument Closer predicts a tendency for a PP-NP-V order, the Principle of Argument Precedence will opt for an NP-PP-V structure, thereby favoring the structure of (13). In this case the two principles would be competing against each other. Therefore when the language has preverbal NP and PP dependents headed by the same verb, the preference for the NP to be adjacent to the head verb will be weaker. In comparison to when the two dependents occur postverbally, this means that the relative order of preverbal NP and PP would be less fixed and more flexible.

The possibly different ordering flexibility between the NP and the PP as well as the potentially contrasting degree of word order freedom between postverbal and preverbal domains across languages laid out here point to a suitable test case for the question raised in this chapter: do syntactic constructions with more ordering variability have a weaker tendency for shorter dependencies? In particular, I chose constructions in which the head verb has a direct object NP dependent and exactly one PP dependent appearing on the same side, with the NP and PP adjacent to each other (e.g. *Kobe praised* $[_{NP}$ *his oldest daughter* $]$ $[_{PP}$ *from the stands* $]$).

6.2.2 Data and preprocessing

I used multilingual corpora from the Universal Dependencies project version 2.5 (UD) (Zeman et al., 2019). I focused on contemporary languages and their treebank data. In every treebank, I searched for VP instances in which the head verb has one direct object NP and one PP dependent adjacent to each other on the same side. Given the UD annotation scheme, I selected verbs with a POS tag of VERB, which denotes verbs that are not auxiliaries. The dependency relation between the nominal head of each NP dependent and the verb was always *obj*, which represents the direct object. The dependency relation between the nominal head of each PP dependent and the verb was always *obl*, which represents oblique. The POS tag of the function head in each PP dependent was ADP and its dependency relation with the nominal head was *case*, which is used for any case-marking element that operates as a separate syntactic unit in UD. In this way PPs

that consisted of only one adposition were not included.

For all instances extracted from each treebank, I first calculated the proportion of instances when the direct object NP appears closer to the head verb as well as the proportion of instances when the PP occurs closer. As the difference in proportion between treebanks for the same language turned out to be small, I combined treebanks for the same language. Languages with fewer than a total of 100 sentences that fit the search criteria (i.e. having a VP with a direct object NP and a PP dependent on the same side, where the two dependents are adjacent to each other) were not included. I coded the word order features as well as the language family and genus of each language following WALS (Dryer and Haspelmath, 2013). After preprocessing, the dataset contained 36 languages in total. Among these languages, twenty-nine ended up being Indo-European, except for Arabic and Hebrew (Afro-Asiatic), Indonesian (Austronesian), Wolof (Niger-Congo), Finish and Estonian (Uralic), and Japanese (Japanese).

6.2.3 Measures

6.2.3.1 Measures for ordering flexibility

To approximate the ordering flexibility of the NP and the PP within each VP instance, I used entropy, a measure of dispersion that reflects the amount of variability within a distribution. As shown in Eq 6.1, X is a binary variable that denotes the relative order of the NP and the PP (e.g. V-NP-PP vs. V-PP-NP; PP-NP-V vs. NP-PP-V); $P(x_i)$ represents the probability of one order. For example, consider a language that has 100 instances where the NP and the PP appear postverbally. Within these instances, 80 appear as V-NP-PP while the other 20 appear as V-PP-NP. The probability of the V-NP-PP structure is then 0.8 (80 / 100 = 0.8), and the probability of the V-PP-NP structure is 0.2 (20 / 100 = 0.2). The ordering freedom for when the NP and the PP are after the head verb in this case is then computed as: $-(0.8 * log_2 0.8 + 0.2 * log_2 0.2) = 0.72$.

$$H(X) = -\sum_{i=1}^{2} P(x_i) log_2(P(x_i)) \tag{6.1}$$

When the probability of one order is 1, which means the probability of the other is 0, the entropy value is 0, indicating there is no flexibility in the order of the two dependents.

On the other hand, when the probability of one order is 0.5, which means the probability of the other is also 0.5, the entropy value is 1, the maximum value for entropy measures. This indicates there is maximum amount of flexibility in the order of the two dependents. In this way, the higher the value of the entropy is (lying between 0 and 1), the more flexible the ordering of the NP and the PP is.

Previous work has applied entropy to measure word order freedom at different levels (Futrell et al., 2015b; Levshina, 2019). Futrell et al. (2015b) adopted conditional entropy to compute word order freedom for a language as a whole. More specifically, they measured ordering flexibility in three aspects: (1) head order entropy, which approximates the ordering freedom of whether syntactic heads appear to the left or the right of their dependents; (2) relative order entropy, which approximates the ordering freedom of the position of a word in a local dependency subtree; (3) relative order entropy of the subject and object, which approximates the ordering freedom of subjects and objects. These measures were all conditioned on specific syntactic features such as the dependency relation type, the POS of the syntactic head and the dependent, and the structure of the dependency subtree. Levshina (2019) used entropy to measure ordering flexibility of particular head-dependent pairs, such as the ordering freedom of determiner and noun, adverbial modifier and adjective, as well as objects and obliques.

My way of approximating flexibility is similar to that in Levshina (2019), except that my focus differs in two aspects. First, although Levshina (2019) noted that there is more flexibility in the order of object and oblique in OV languages and flexible languages, an observation that also aligns with that from Hawkins (2014), the results did not distinguish the difference of ordering freedom between preverbal and postverbal domains. Secondly, Levshina (2019) took all oblique phrase types into account, which do not only include PPs based on the UD annotationa scheme (e.g. certain nominal modifiers that denote temporal or locational relations are also annotated as obliques). In comparison, the experiment settings in this chapter are more controlled and fine-grained, since oblique phrases of different syntactic categories might behave differently.

Using entropy as an approximation for the variability in the order of the NP and the

PP, I estimated 95% confidence intervals for ordering flexibility with bootstrapping (Efron, 1979) for 1,000,000 iterations. The basic procedures for computation were as follows. (1) Given a dataset of n VP instances in total, I randomly drew a sample of n instances from the dataset with replacement. In this case the same instance from the dataset could appear more than once; (2) within the sample, I measured the ordering flexibility of the NP and the PP with entropy; (3) I repeated step (1) and (2) for 1,000,000 iterations, which resulted in an empirical resampling distribution of the entropy value from each iteration; (4) I calculated the mean and 95% confidence interval of the distribution derived from step (3).

6.2.3.2 Measures for DLM

Here I measured the strength of DLM as the predictive power of dependency length in the relative order of the NP and the PP within a particular syntactic construction. To more accurately compute the effect of dependency length, I used mixed-effect logistic regression modeling, including two automatically measurable factors as fixed effects: argument status and the pronominality of both the NP and the PP, and the head verb within each instance as a random effect. I considered the direct object NP to be more argument-like than the PP in each VP instance. The modeling procedures were performed in the same way as in Chapter 4. Estimates for 95% confidence intervals of the coefficient of each factor were derived from 10,000-fold cross-validations. If the syntactic construction has a preference for shorter dependencies, the coefficient of dependency length in predicting the order of the NP and the PP in that construction will be significantly larger than 0.

6.2.3.3 Measures for correlation between ordering flexibility and DLM

I used Spearman's ρ, a nonparametric measure to approximate the correlation between ordering flexibility and the extent of DLM. The value of Spearman's ρ assesses whether there is any monotonic relationship (not necessarily linear) between two variables in question. The score of Spearman's ρ ranges from -1 to 1, with -1 indicating a perfect negative correlation while 1 representing a perfect positive correlation. When the value of ρ is 0, this means there is no correlation at all. Again, if there is longer dependency in structures with higher degrees of word order freedom, there will be an overall negative correlation

between the degree of ordering flexibility and the extent for DLM.

6.3 Results

Overall the dataset exhibits three ordering patterns:

- one for languages with NP and PP after the head verb (e.g. Wolof; Finnish);
- one for languages with NP and PP before the head verb (e.g. Persian; Hindi);
- one for languages with NP and PP both after and before the head verb (e.g. Czech; Estonian).

6.3.1 Languages with postverbal orderings

Of all the languages, twenty have the NP and the PP dependents both appearing after the head verb. As shown in Figure 6.1, most of these languages are Germanic or Romance languages. In particular, the four Germanic languages, Danish, English, Norwegian and Swedish all have a relatively low degree of word order freedom, in contrast to that in the Romance languages. Among these languages, Finnish has postpositional PP whereas the rest all have prepositional PP. Nevertheless, ordering flexibility does not seem to depend on the headedness of the PP. The degree of word order freedom in Finnish (0.85) is comparable to languages with prepositional PP, such as Galician (0.86) and Greek (0.82).

All of these languages exhibit a strong tendency for shorter dependencies. On the other hand, I did not find a negative correlation or any correlation between ordering flexibility and the effect for dependency length (Spearman's $\rho = 0.08$). This means that among these languages, when they have more ordering flexibility, they do not tend to have a weaker extent for DLM.

6.3.2 Languages with preverbal orderings

In five languages from my dataset, the NP and the PP dependents both occur before the head verb when they are on the same side. As shown in Figure 6.2, their relative order appears to be more flexible in comparison to when the two dependents are postverbal. While admitting the relatively small size of this group of languages, there does not seem to

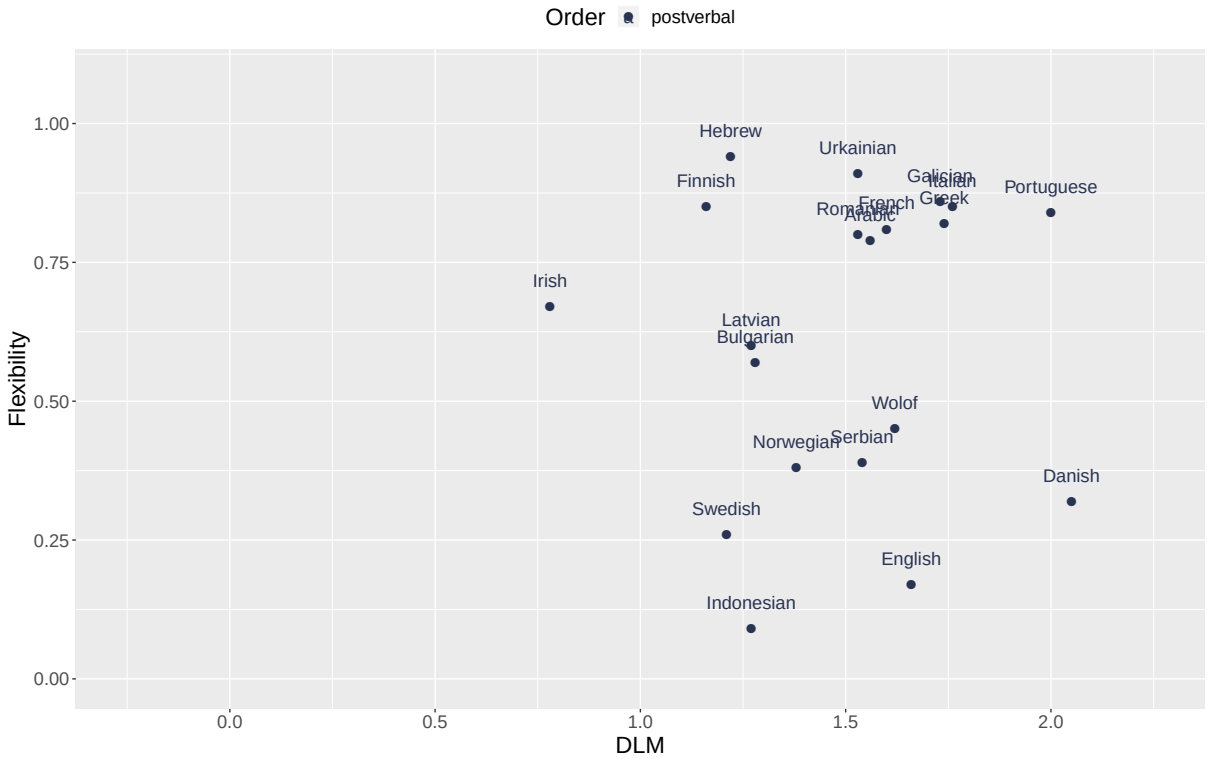


Figure 6.1: Scatterplot of ordering flexibility against DLM in languages with postverbal NP and PP. Average ordering flexibility is 0.62 and average DLM is 1.49.

be a strong discrepancy in ordering freedom concerning whether the PP is prepositional or postpositional. Three of the five languages have postpositional PP: Japanese, and two Indic languages, Urdu and Hindi, which are all rigid OV. The flexibility for these three languages is comparable to each other. In contrast, Afrikaans and Persian have prepositional PP and they are both considered non-rigid OV. The ordering flexibility in Afrikaans is comparable to the three rigid OV languages, though Persian seems to have a much higher degree of word order freedom.

On the other hand, the effect for dependency length is much weaker for these languages on average than for the languages of Figure 6.1. There exists a strong negative correlation between ordering flexibility and the extent of DLM (Spearman's $\rho = -0.90$), suggesting that for these languages, structures with more ordering flexibility have longer dependencies.

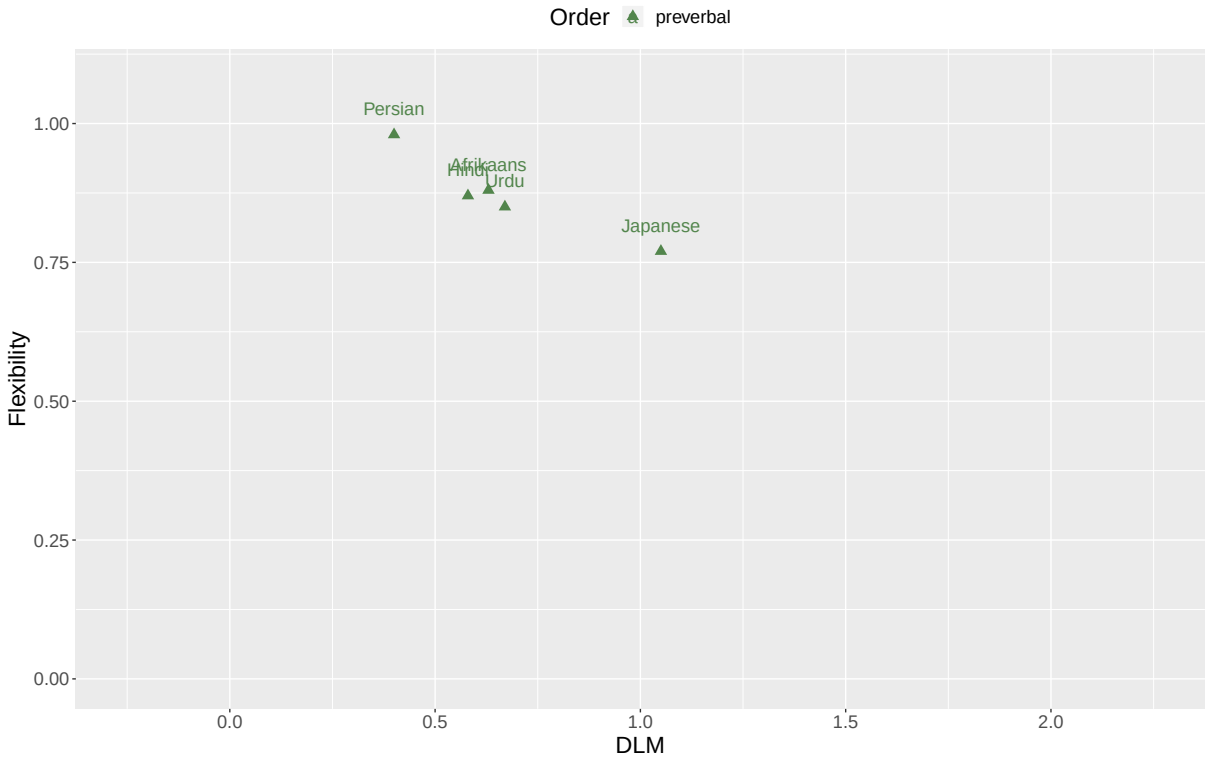


Figure 6.2: Scatterplot of ordering flexibility against DLM in languages with preverbal NP and PP. Average ordering flexibility is 0.87 and average DLM is 0.67.

6.3.3 Languages with both postverbal and preverbal orderings

Now consider languages with mixed ordering patterns, that is, languages in which the NP and the PP can appear both after and before the verb, as in Figure 6.3. These languages mainly fall into three subtypes: Germanic, Romance and Slavic. Overall most of these languages demonstrate great variability in the order of the two dependents in postverbal as well as in preverbal domains. On average, the ordering flexibility of the two dependents when they are after the verb is similar to when they occur before the verb; it is also comparable to the languages with only preverbal orderings in Figure 6.2, and is much higher than in languages where the two dependents only occur after the verb (Figure 6.1).

In addition, the ordering freedom in these languages does not seem to be constrained by whether the PP is prepositional or postpositional either. Estonian is the only language with postpositional PP in these languages, and its ordering flexibility is comparable to that

in certain languages with prepositional PP such as Czech or German in both preverbal and postveral domains. On the other hand, for these languages, there appears to be a strong negative correlation between ordering flexibility and the extent for DLM when the two dependents appear after (Spearman's $\rho = -0.52$) or before (Spearman's $\rho = -0.67$) the head verb.

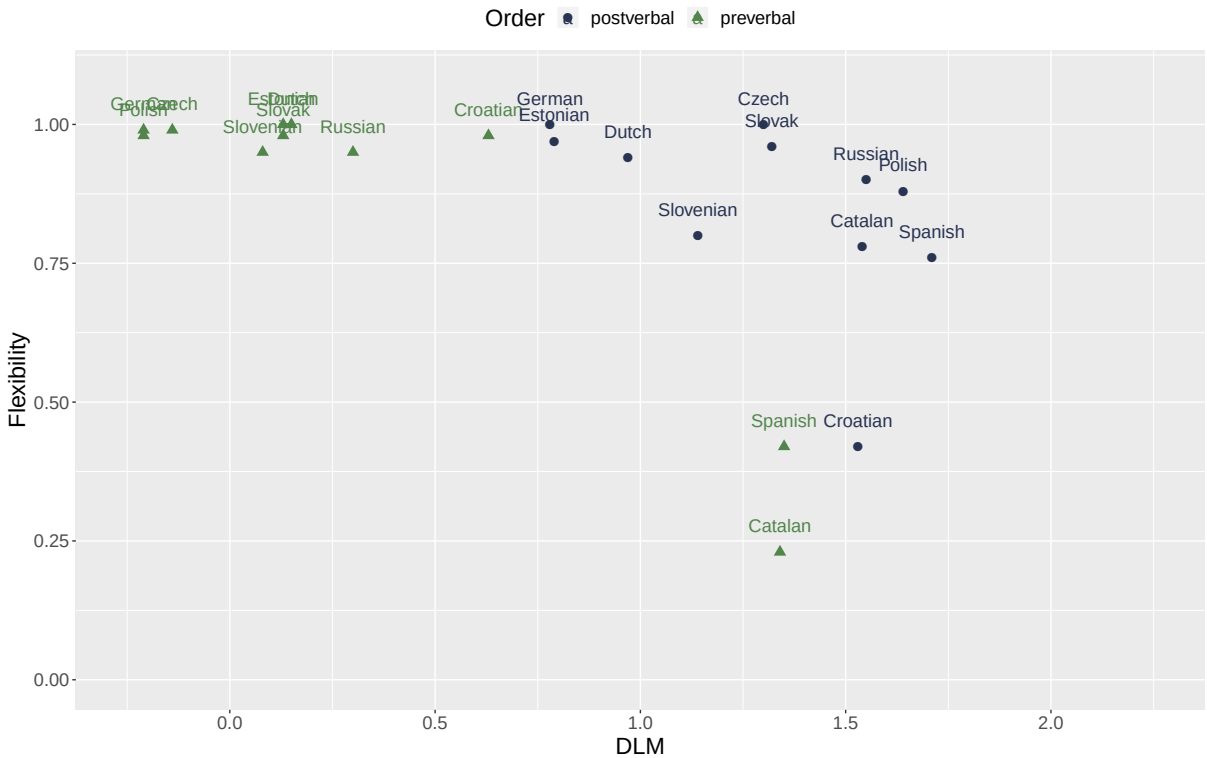


Figure 6.3: Scatterplot of ordering flexibility against DLM in languages with mixed orderings. In postverbal domains, average ordering flexibility is 0.86 and average DLM is 1.30. In preverbal domains, average ordering flexibility is 0.85 and average DLM is 0.32.

6.3.4 An overall look at all languages

Combining the results from Figure 6.1, Figure 6.2 and Figure 6.3, in general there is more ordering flexibility when the NP and the PP dependents occur in preverbal domains (average flexibility is 0.86), in comparison to when they occur in postverbal domains (average flexibility is 0.76). The least flexible languages appear to be those that are predominantly prepositional, such as English and Indonesian. On the other hand, the most flexible languages are mainly those where the NP and the PP dependents can appear

both after and before the head verb, such as German and Slovak.

Similar to findings from Chapter 3 and Chapter 5, there is a much stronger preference for shorter dependencies when the NP and the PP are after the head verb, in contrast to when they are before the head verb. This distinction is most obvious in languages with mixed ordering patterns, where the two dependents can occur both postverbally and preverbally. Overall there seems to be a strong tendency for a weaker extent of DLM as ordering flexibility increases (Spearman's $\rho = -0.58$).

6.4 Discussion

In this chapter, I have asked whether structures with more word order freedom have a weaker preference for shorter dependencies crosslinguistically. Using syntactic constructions with a direct object NP and a PP dependent adjacent to each other on the same side as a simple test case, I have shown that the answer is yes. Overall, I have presented evidence for a negative correlation between ordering variability and the extent of DLM: as the degree of word order variability increases, there is a weaker tendency for shorter dependencies. In particular, there is on average more ordering flexibility in preverbal domains in contrast to postverbal domains. Aligning with findings from Chapter 3 and Chapter 5, I demonstrated that the effect for dependency length is much weaker in preverbal orderings than in postverbal orderings.

Although pronominality is not the focus in this study, in general there is a strong ordering effect for pronominality across languages. The preference for the pronominal phrase to occur first, whether it is the NP or the PP, appears to hold both after and before the head verb. Following the general pattern that pronominal elements tend to be more "given" than nonpronominal items, the preferences demonstrated here support the prediction made by the Principle of Given-before-new (Bock, 1986; Haviland and Clark, 1974; Ferreira and Lowder, 2016; Christianson and Ferreira, 2005; Ferreira and Yoshita, 2003).

For future studies, it would be worthwhile to see whether results from behavioral experiments addressing the same question would yield similar findings. Taking a gradient

approach, Namboodiripad (2019) adopted acceptability judgement tasks and used acceptability scores to measure degree of ordering freedom in English and Malayalam. A similar methodology could be applied to any other syntactic constructions crosslinguistically in order to bring more insights into the word order features of different languages from the angle of their flexibility.

Chapter 7

Frequency-dependent Regularization in Constituent Ordering Preferences

7.1 Introduction

The experiments from Chapter 3 to Chapter 6 have demonstrated that when grammatical alternatives exist for a given syntactic construction with flexible constituent orderings, the relative orders of the constituents are governed by a variety of abstract constraints, such as dependency length (Futrell et al., 2015a), semantic closeness (Hawkins, 1999) and discourse information (Prat-Sala and Branigan, 2000). The respective roles of these abstract constraints have been the focus so far in order to explain ordering preferences across languages.

On the other hand, previous studies have noted that constituent orderings are also influenced by item-specific knowledge of the particular syntactic construction in question. This might in turn overrule the effects of other abstract factors and lead language users to have idiosyncratic ordering preferences (Morgan and Levy, 2016a, 2015; Ferreira, 1994). For example, the ordering preference for binomial expressions in English (X and Y) is largely affected by lexical, semantic and phonological properties of the words in the binomials, such as that the shorter word appears first; the final syllable of the second word should not be stressed. Given these different abstract factors, within the binomial construction type that includes both *safe and sound* and *sound and safe*, there might not be a sharp discrepancy in the preference strength or extremity for the two alternatives

since, for instance, although *safe* has fewer letters than *sound*, *sound* is chronologically more frequent than *safe*. Nevertheless, language users prefer *safe and sound* due to more frequent usage of this particular order. In other words, *safe and sound* is consistently preferred over *sound and safe* not because of the abstract constraints, but rather because language users have item-specific knowledge of this binomial type.

Within the context of ordering preferences, the consistent preference for one structural variant among all syntactic alternatives (e.g. *safe and sound* is regularly more preferred to *sound and safe*) is known as *regularization*, a well-known phenomenon in statistical learning (Reali and Griffiths, 2009). In the linguistic domain in general, *regularization* refers to the tendency to make language structures more systematic and fixed, which minimizes the extent of variation in language usage. [1] Here specifically, regularization refers to the phenomenon that when there is variation in the input, language users will preferentially reproduce the most frequent alternative that they have encountered. [2] For example, if a speaker hears the binomial type that includes *safe and sound* and *sound and safe*, the former will be more preferred in production due to its higher overall frequency in the input. Previous experiments in language learning and production have shown that both adults and children tend to regularize their output given the input (Hudson Kam and Newport, 2005), which can potentially explain why truly unpredictable or free linguistic variation is rare.

Nevertheless, the pressure to regularize contradicts the dominant view from research situated in rational language processing (Levy, 2008). This line of work posits with mounting evidence that language users are sensitive to the probabilistic distributions of different linguistic structures. Therefore they should perform *probability matching* rather than regularization. For a given structural type, they will reproduce all alternatives such that the ratios for these alternatives match their original probability in the input. For instance, if a speaker has encountered *safe and sound* 80 times and *sound and safe* 20 times (a 4:1 ratio), in their production the ratio for these two variants will approximate

[1] The notion of *regularization* here differs from morphological regularity which refers to certain linguistic items abiding by compositional rules when going through morphological processes.

[2] Similar predictions are also made by the concept of *entrenchment* in the literature of cognitive grammar (Langacker, 1987).

4:1 as well. In this case, instead of minimizing variation, language users are likely to exhibit behaviors that maintain the same or similar distributions to the given input.

Morgan and Levy (2016b) demonstrated regularization in ordering preferences for binomial expressions in English. More importantly, they showed that the extent of regularization is frequency-dependent: the higher the overall frequency of the binomial type, the more regularized and extreme is the preference that language users have for one alternative over the other (*radio and television* > *television and radio*; *salt and pepper* >>> *pepper and salt*). Here overall frequency refers to the unordered frequency. For example, the overall frequency of the binomial type that contains *salt and pepper* as well as *pepper and salt* sums up both the frequency of the former and the frequency of the latter. They further demonstrated that it is possible to account for this frequency-dependent regularization bias with computational modeling of language change. Specifically, they adopted Iterated Learning Models which can simulate how language changes over generations (Kirby et al., 2014).

As fruitful as these previous findings are, most studies which have demonstrated a regularization bias have focused on the learning and production of individual words or small phrases (Hudson Kam and Newport, 2005), while exploration of regularization in syntactic constructions at a higher level has been lacking. Additionally, previous research along these lines has mainly used behavioral experimental data from artificial learning (Ferdinand et al., 2019; Culbertson and Smolensky, 2012) rather than naturalistic corpora. Thus in general, whether language users tend to perform probability matching or regularization when reproducing structural variants and under what contextual conditions remain far from clear.

This chapter makes a contribution towards this gap. Following Morgan and Levy (2016b), I investigated the role of verb idiosyncrasy in constituent ordering preferences for abstract syntactic constructions above the word level. Leveraging large-scale corpus data, I addressed two questions. First, does the same frequency-dependent regularization found at the word level for binomials in English also operate on more complex syntactic levels? Secondly, how does this frequency-dependent regularization bias emerge diachronically?

In comparison to binomial expressions, the regularization pattern in syntactic structures above the word level might be different. The length of binomials is relatively short (in Morgan and Levy (2016b) all expressions have a length of 3 words) and it mostly involves orderings of two words (e.g. whether to put *safe* before *sound* or vice versa). By contrast, larger constituents in syntactic constructions above the word level tend to be much longer. And deciding the relative order of larger constituents potentially involves more processing effort in comparison to binomials, therefore, which may lead to a stronger regularization bias (Ferdinand et al., 2019).

On the other hand, given that more abstract structures possibly contain more syntactic complexity than binomials, when ordering larger constituents, it might be the case that there are more lexical and structural constraints at different linguistic levels that should be taken into consideration. Accordingly, even if verbs have idiosyncratic preferences, they may not be exerting these to a significant effect. In this case there might be less regularization in syntactic constructions above the word level.

7.2 Experiments

7.2.1 The test case

To address whether frequency-dependent regularization exists in more abstract syntactic constructions as well as the origin of the regularization bias, I used the dative construction (Bresnan, 2007; Bresnan and Ford, 2010; Yi et al., 2019) in English as the test case.

There is no definitive and concrete criterion to judge whether a verb is a dative verb or not. Based on work by Bresnan et al. (2007), this chapter treats a verb as a dative verb if it has been noted to belong to either the dative alternation class (e.g. (1) and (2) with the recipient *mama*) or the benefactive alternation class (e.g. (3) and (4) with the beneficiary *papa*) in the seminal work of Levin (1993). [3] In this way different dative types are distinguished based on their head verbs.

(1) I **sent** [$_{NP}$ mama] [$_{NP}$ some flowers].

[3] Verbs were taken directly from `http://www-personal.umich.edu/~jlawler/levin.verbs`. There are 336 verbs listed in the dative alternation class and 177 in the benefactive alternation class, with 23 verbs overlapped in both classes.

(2) I **sent** [$_{NP}$ some flowers] [$_{PP}$ to mama].

(3) I **poured** [$_{NP}$ papa] [$_{NP}$ some tea].

(4) I **poured** [$_{NP}$ some tea] [$_{PP}$ for papa].

Using the dative construction, I examined whether there is a relationship between regularization/preference extremity and the overall frequency of each dative type. I further investigated the origin of the regularization bias with computational modeling of language change, exploring in particular whether Iterated Learning Models can predict frequency-dependent regularization.

7.2.2 Overall construction frequency

To address whether frequency-dependent regularization exists in the dative construction, I needed to measure both the overall frequency of each dative type, as well as the preference extremity within each dative type. Estimates of overall frequency require large amount of training data, for which I used UD-style corpus in English taken from the raw data of the CoNLL 2017 Shared Task on multilingual parsing (Ginter et al., 2017). This corpus, automatically parsed with UDPipe (Straka and Straková, 2017), has a total of around 9 billion tokens and consists of web page data from both Common Crawl and Wikipedia. One caveat with this corpus is that it is not labeled with gold-standard semantic roles. This makes it difficult to search more definitively for sentences that contain a recipient or a beneficiary, which is presumably required in a dative construction.

As an alternative approximation, I searched for VP instances where the head verb belongs to the list of dative verbs as defined in Section 7.2.1 and appears in either the double object structure (V-NP-NP) or the prepositional object structure (V-NP-PP). For double object structures, I extracted sentences in which the head verb takes one direct object and one indirect object (in a V-NP-NP structure, the first NP is the indirect object and the second is the direct object given the corpus annotation). For prepositional object structures, I extracted sentences in which the head verb takes one direct object and one PP oblique that immediately follows the direct object; the functional head of the PP was always *to* if the verb belongs to the dative alternation class, and *for* if the verb belongs to the benefactive alternation class. The dependency relation between the nominal head

of the direct object and the verb was always *obj*; the dependency relation between the nominal head of the indirect object and the verb was always *iobj*; and the dependency relation between the nominal head of the PP and the verb was always *obl*.

It is true that not every VP extracted this way has a grammatical alternative. In particular, if a verb appears in the V-NP-PP structure, its V-NP-NP alternative might not be deemed grammatical under certain circumstances. As an illustration, consider the following sentences drawn from the corpus search (slightly adapted).

(5) ...I will **show** [$_{NP}$ it] [$_{PP}$ to him]...

(6) *...I will **show** [$_{NP}$ him] [$_{NP}$ it]..

(7) ...a private company is going to **fly** [$_{NP}$ a crew] [$_{PP}$ to MARS]...

(8) *...a private company is going to **fly** [$_{NP}$ MARS] [$_{NP}$ a crew]...

(9) We've **created** [$_{NP}$ this conference] [$_{PP}$ for specific dealers]...

(10) *We've **created** [$_{NP}$ specific dealers] [$_{NP}$ this conference]...

(11) ...I can **save** [$_{NP}$ time] [$_{PP}$ for other things]...

(12) *...I can **save** [$_{NP}$ other things] [$_{NP}$ time]...

(13) ...she **made** [$_{NP}$ this] [$_{PP}$ for dinner]...

(14) *...she **made** [$_{NP}$ dinner] [$_{NP}$ this]...

Nevertheless, this does not necessarily pose an issue with the questions to be addressed in this chapter (i.e. frequency-dependent regularization in the dative construction). The experiments here focus on the idea that if a dative verb occurs in a prepositional object order, the PP can be interpreted as having the meaning of a beneficiary or a recipient in at least some contexts (perhaps more creatively for some PPs than others; e.g. *explain this to me* vs. *explain me this* (Goldberg, 2019)). Thereby the structure has the potential to occur in a double object order, regardless of whether it is always deemed grammatical.

Another motivation is that the notion of grammaticality varies among different language users under different scenarios. One crucial point noted in previous experiments (Bresnan et al., 2007) is that not every alternation headed by a verb has to have a grammatical alternative, because whether there is a grammatical syntactic alternation is constrained

by abstract factors such as dependency length and animacy. To illustrate this, consider the following examples. The verb *give* is perhaps one of the most typical dative verb, yet many consider (16) as ungrammatical because Budapest is not an animate object, and accordingly it can not serve as the recipient of the action.

(15) She gave the draft to him.

(16) She gave the draft to Budapest.

On the other hand, the grammaticality of (16) might be regarded differently in a particular context. The sentence can be perfectly fine if it is used as a metonym (e.g. referring to the publisher office located in Budapest) (Hovav and Levin, 2008).

As mentioned before, different dative types are distinguished by different head verbs. Therefore sentences with the same head verb are considered as belonging to the same dative type. Following this criterion, I grouped all extracted instances based on their head verbs into distinct dative types, regardless of whether the instance is realized as a prepositional object structure or a double object structure. Then I calculated the overall frequency of each type. Dative types with an overall frequency lower than 1000 were removed to ensure that there is enough data to reliably estimate the frequency of different alternations. After preprocessing, the dataset contains 225 unique dative types (225 unique head verbs) with a total of 9.3 million dative instances. Among these, 4.0 million appear in the double object structure while 5.3 million appear in the prepositional object structure.

7.2.3 Preference extremity

I approximated the preference extremity within each dative type as the probability for the more preferred structure between the two alternations of each dative type. Since my goal was to estimate the role of verb idiosyncrasy in the preference extremity, and the ordering preferences of the dative construction are known to be constrained by abstract factors such as phrasal length and pronominality (Bresnan et al., 2007; Bresnan, 2007), the effects of these factors have to be excluded in order to more accurately quantify the influence of verb idiosyncrasy.

To do this, I first fit a mixed-effect logistic regression model to predict the prepositional object structure in the dataset. I included verb as a random effect and included fixed effects for three automatically measurable factors: definiteness, pronominality and phrasal length. It would be more ideal if the model contained additional factors that have been tested in previous dative alternation studies such as animacy and the semantic class of the head verb within contexts. However, those factors require manual coding, which is not realistic given the current experimental settings.

Since the focus was on how specific verbs affect ordering preference extremity, I estimated the effect of verb bias (Wasow and Arnold, 2003; Stallings et al., 1998) as the probability of a sentence being realized as the prepositional object structure predicted from just the random effect of the verb (eliminating the contributions from the fixed effects). Specifically, let V be the random effect intercept of a particular verb derived from the regression model. The probability of this verb being realized in the V-NP-PP structure is then calculated as follows.

$$\frac{1}{1 + exp(-1 * V)} \tag{7.1}$$

Given each dative type, a probability value larger than 0.50 indicates the prepositional object structure is preferred over the double object order, and the preference extremity is the same as the probability value. When the probability is lower than 0.50, this indicates a stronger preference for the double object order, and the preference extremity is computed as the absolute difference between this probability value and 1.

After collecting both the overall frequency and preference extremity of each dative type, I fit a linear regression with the former being the predictor and the latter being the outcome variable in order to evaluate the significance of overall frequency.

7.2.4 Iterated Learning Models

In this section I took up the second question raised at the beginning: how does frequency-dependent regularization occur in the first place? To address this, I borrowed Iterated Learning Models (hereafter ILM), which are computational models that simulate language change. In what follows, I introduced both standard 2-alternative ILMs and the

augmented ILMs from Morgan and Levy (2016b). The models employed in this chapter followed Morgan and Levy (2016b). Then I described the simulation procedures of the models for the dative corpus data.

7.2.4.1 Standard ILMs

Iterated learning (Kirby, 2001; Kirby and Hurford, 2002) has gained wide popularity over recent years as an approach to study how language evolves through cultural transmission. The crucial insight of this methodology is that language structures are transmitted culturally via language users learning linguistic constructions from others' usage patterns of the same constructions. Meanwhile during each stage of transmission and learning, language learners can impose their own biases on the usage of the constructions as well, which in turn shapes language structures.

ILMs computationally simulate this learning process, where the output of the previous learner is fed as the input to the next learner and this process proceeds in an iterative fashion. If the tendency to regularize emerges from reproduction of structural alternatives based on the input, which is also a process that continues iteratively, ILMs serve as an ideal tool to account for regularization.

The learning process of ILMs is a process of Bayesian inference. For each dative type, the basic idea for the learning process is as follows. The first learner at the beginning of the learning process is initialized with a hypothesis θ_1, which is the probability of the dative type appearing as the prepositional object structure. Then the learner is going to produce a total of N instances for that particular dative type, with there being x in the prepositional object order, and $N - x$ in the double object order. Given the output from the first learner, together with the prior knowledge of the dative type being in the prepositional object order, the second learner will perform Bayesian inference and infer a hypothesis θ_2, which also represents the probability of the dative type appearing in the prepositional object structure, and possibly has a different value compared to that of θ_1. The second leaner continues to generate new data given θ_2, and the following learners repeat the same procedure.

Mathematically, the prior of the dative type being in the prepositional object order (the

probability of θ in general) is expressed as the beta distribution (B) with two parameters: μ and ν. The former defines the mean of the distribution while the latter determines the width or the concentration of the distribution.

$$P(\theta) = \frac{\theta^{\mu\nu-1}(1-\theta)^{(1-\mu)\nu-1}}{B(\mu\nu, (1-\mu)\nu)} \tag{7.2}$$

As for interpretation, μ represents the ordering preference for a given dative construction type. The learners in all generations are set to have the same μ in the learning process of each dative type. As I assumed learners have no innate knowledge of which structure will be more preferred and what their relative probabilities should be, I assigned μ with a value of 0.5. This means that the learner believes the two alternative structures have equal probability of occurrence.

On the other hand, ν reflects the confidence in the prior probability, with a higher value representing the learner's greater confidence about their prior knowledge, and a lower value representing their lesser confidence. Different from μ, ν is a free parameter. To demonstrate how the value of ν affects the shape of the prior distribution, I simulated 10,000 data points following the beta distribution with different values of ν. As shown in Figure 7.1, when μ is 0.5, a higher value of ν indicates that the prior distribution centers around 0.5. This means that most of the time the learner believes the prepositional object structure and the double object structure will appear for roughly equal proportions of times, which corresponds to more structural variation. A lower value of ν denotes that the prior distribution is more scattered. This means that most of the time the learner believes within each dative type, one alternation is more preferred than the other, which leads to more regularization. When the value of ν is 2, this conforms to the uniform distribution, which means that the learner does not have any strong belief regarding which alternation will be more preferred for a dative type.

Given a learner's hypothesis θ, new data are generated following the binomial distribution.

$$P(x|\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{7.3}$$

The next learner applies Bayes rule to calculate a posterior distribution over all hypotheses

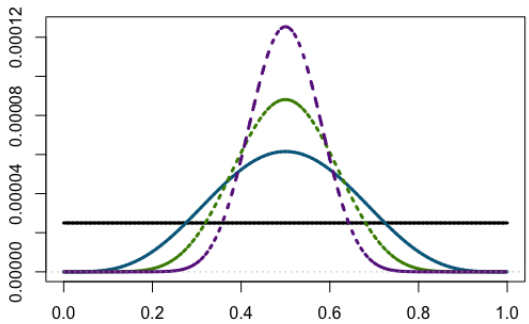


Figure 7.1: Plot of prior distribution when μ is 0.50 with different ν values: $\nu = 2$ (black, which is the uniform distribution); $\nu = 10$ (blue); $\nu = 20$ (green); $\nu = 40$ (purple).

of the dative type being realized as the prepositional object order.

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \tag{7.4}$$

Since the beta distribution is the conjugate prior of the binomial distribution, the posterior follows the same distribution as the prior. In other words, the posterior distribution also corresponds to the beta distribution. The learner then samples a hypothesis from this posterior distribution and generates new data.

7.2.5 ILMs from Morgan and Levy (2016b)

Though standard ILMs are able to show a general tendency for regularization, they cannot capture the relationship between the preference extremity and the frequency of an expression. To solve this, Morgan and Levy (2016b) augmented standard ILMs at the stage of data generation. Within each generation of an augmented model, the learner applies a regularization function (Eq. 7.5) [4] with a free bias parameter R to update θ and form a new hypothesis θ'. The regularization function itself is frequency-independent, in the sense that the value for R holds in each stage for all dative types. When R equals 1, as presented in Figure 7.2, the new hypothesis θ' is the same as the original θ. On the other hand, as the value of R increases, this corresponds to more and more pressure to

[4]The regularization function here is different from the one in Morgan and Levy (2016b) and can be applied to constructions with more than two alternations.

regularize. The model then generates data based on θ'.

$$\theta' = \frac{\theta^R}{\theta^R + (1-\theta)^R} \tag{7.5}$$

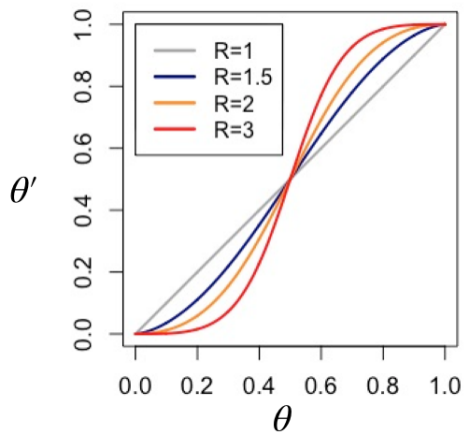


Figure 7.2: Plot of θ' against θ given different values of R: $R = 1$ (grey; $\theta' = \theta$), $R = 1.5$ (blue), $R = 2$ (orange), $R = 3$ (red).

In Morgan and Levy (2016b), the incorporation of the frequency-independent regularization bias in the augmented ILMs has led to the emergence of frequency-dependent regularization and the models were able to predict the preference extremity for binomials as observed in corpus data.

7.2.6 Simulating the dative corpus data

To predict the frequency-dependent regularization in the dative construction, I followed Morgan and Levy (2016b) and introduced a frequency-independent regularization function with a free parameter R to modulate original θ to θ' at each data generation stage. In the simulation process, I set N of each dative construction type to be an approximate for the number of times a college student who is a native speaker of English has been exposed to that particular dative type. The estimate for the lifetime linguistic exposure of a college-age native English speaker is around 300 million words in total (Levy et al., 2012). For each dative type, I ran 80 chains of learners for 1500 generations. This is not to suggest that realistically the process of language learning and production has continued for 1500 generations, but rather in order for the model's learning process to

reach the stationary distribution (when the posterior distribution converges to the prior distribution and stays stationary as each stage of learning continues (Kirby et al., 2007)). Within each chain, the initial hypothesis θ_1 is assigned a value of 0.50. For hypothesis updating, I experimented with a series of different values for the two free parameters ν and R ($\nu = \{2, 3, 4, 5, 6\}$; $R = \{1, 1.1, 1.5, 1.8, 2, 2.1, 2.5, 2.8, 3, 3.1, 3.5, 3.8, 4\}$), resulting in a total of 65 models. I collected the θ' from the final generation of each chain. A θ' value higher than 0.50 indicates a preference for the prepositional object structure over the double object structure, and the predicted preference extremity by the model is the same as θ'. If θ' is smaller than 0.50, the double object structure is the more preferred order between the two alternatives and the predicted preference strength is measured as 1 - θ'.

7.3 Results

7.3.1 Existence of frequency-dependent regularization

As presented in Figure 7.3, frequency-dependent regularization does exist in the dative construction ($\beta = 0.03, p < 0.05$). The higher the overall frequency of the dative type is (denoted by the head verb), the stronger the preference extremity is for one alternation over the other. Among the most extreme cases are the typical dative verbs such as *give*, *ask*, *call*, which all favor double object order as predicted by the logistic regression model.

7.3.2 Accounting for frequency-dependent regularization

After validating that frequency-dependent regularization exists in the dative construction, I evaluated the predictions of the ILMs for the corpus data of the dative construction. If ILM is able to account for regularization, as the overall frequency of the construction type increases, the value for the predicted preference extremity should increase as well. Results from Figure 7.4 corroborate findings in Figure 7.3, showing that the model can predict frequency-dependent regularization in the dative construction with combinations of appropriate values for ν and R, though to different extents (Table 7.1).

It appears that a comparable pattern to that in Figure 7.3 can already be derived when when ν equals 2. This lends support to the earlier motivation of initializing the

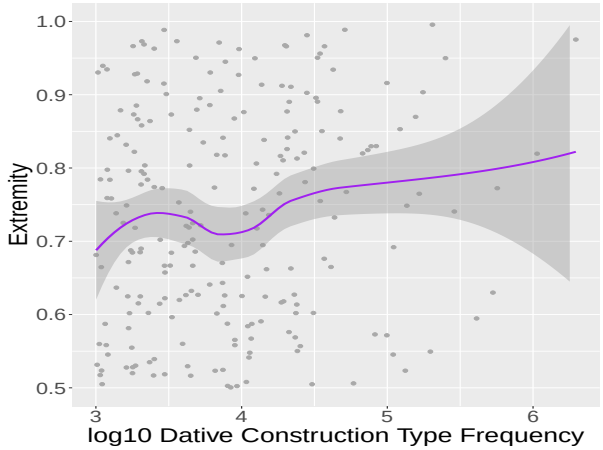


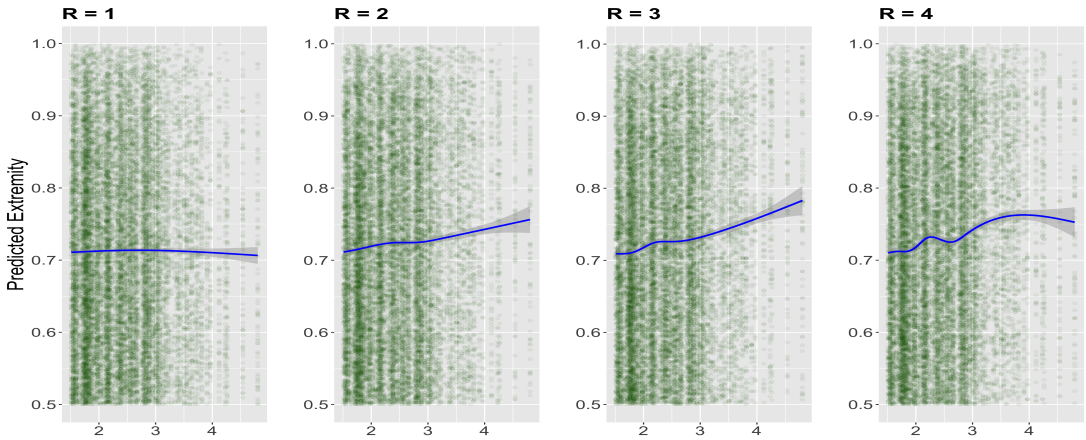
Figure 7.3: Plot of preference extremity against log10 overall construction frequency for 225 dative construction types. Overall frequency is estimated from corpus data for English from Ginter et al. (2017).

prior probability as 0.50, since a beta distribution with μ of 0.50 and ν of 2 is the uniform distribution, which means that the prior is truly uninformative, i.e. learners have no innate knowledge of which structure is more preferred. When the ν value is held constant, the pressure to regularize is stronger as the value of R increases.

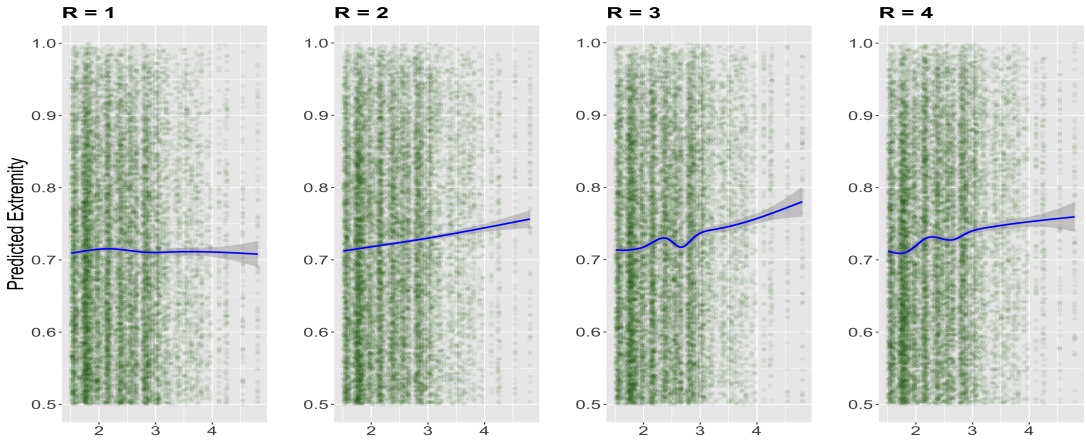
The observations here differ from Morgan and Levy (2016b) in one respect. They demonstrated that regularization in binomial expressions in English already emerges in their models when R is as low as 1.1, yet with much larger values for ν ($\nu = \{10, 15, 20\}$). Recall that a lower value for R as well as a higher value for ν both correspond to a weaker regularization bias. This means that the extent of regularization is stronger in the dative construction than in binomials. I return to this point in Section 7.4.

	$\nu = 2$	$\nu = 3$	$\nu = 4$
$R = 1$	$(\beta = 0.00, p > 0.05)$	$(\beta = 0.00, p = 0.60)$	$(\beta = 0.00, p = 0.60)$
$R = 2$	$(\beta = 0.01, p < 0.01)$	$(\beta = 0.01, p < 0.01)$	$(\beta = 0.01, p < 0.01)$
$R = 3$	$(\beta = 0.02, p < 0.01)$	$(\beta = 0.02, p < 0.01)$	$(\beta = 0.01, p < 0.01)$
$R = 4$	$(\beta = 0.02, p < 0.01)$	$(\beta = 0.02, p < 0.01)$	$(\beta = 0.02, p < 0.01)$

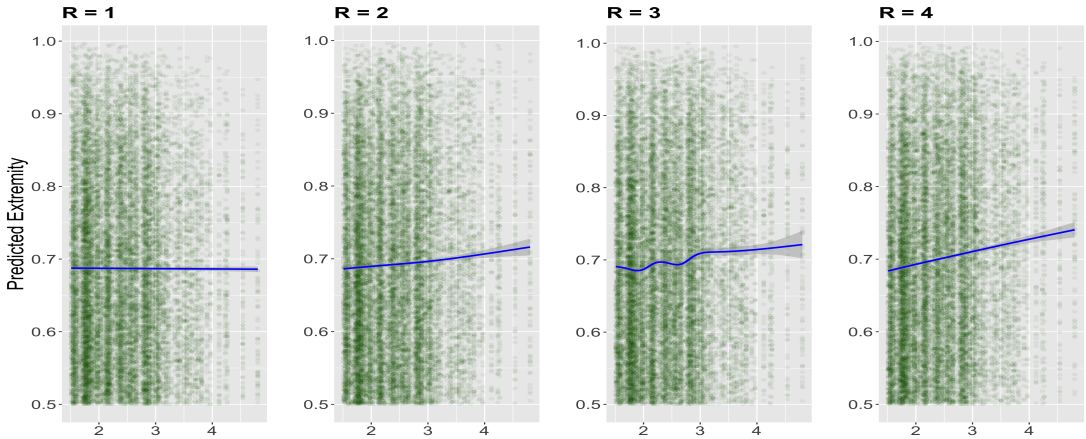
Table 7.1: Linear regression (predicting preference extremity as a function of overall frequency) results for subgraphs in Figure 7.4.



(a) $\nu = 2$



(b) $\nu = 3$



(c) $\nu = 4$

Figure 7.4: Selected plots of predicted extremity values from ILM with different values for R and ν.

7.4 Discussion

Using the dative construction in English as the test case, I have demonstrated that there is frequency-dependent regularization in constituent ordering preferences in abstract syntactic constructions above the word level. The more frequent the construction type is, governed by the head verb, the more polarized a preference language users have for one syntactic variant over the other. In addition, the second question I have addressed concerns the origin of frequency-dependent regularization. Recall that while standard ILMs are not able to account for frequency-dependent regularization, the results have shown that when combined with a frequency-independent regularization bias, the augmented models are able to predict the observed regularization patterns in the dative construction. This indicates that just language processing alone is not enough to yield frequency-dependent regularization, but rather this pattern arises from the continuous interactions between language production and the process of cultural transmission and language learning.

In contrast to the results in this chapter, previous experiments on verb idiosyncrasy have mainly focused on comprehension tasks rather than the production (MacDonald, 1994; Garnsey et al., 1997). They have demonstrated that comprehenders perform probability matching and that the probabilistic information of verb subcategorization frames is able to predict processing behaviors. For example, the verb *suggest* has a stronger preference for taking a sentential complement (e.g. *We* ***suggest*** *that it is time to investigate verb bias.*), rather than having a direct object (e.g. *The reviewers* ***suggest*** *more examples.*). Accordingly, a sentence in which *suggest* is followed by a sentential complement is proportionally easier to process compared to one where *suggest* takes a direct object.

Overall, all these findings indicate that a cohesive and complete account for language structures and processing patterns should incorporate item-specific knowledge along with abstract factors. Indeed, Morgan and Levy (2015) have shown that the model which has the best performance in predicting the distribution of preference polarization for binomial expressions in English is the one that takes into account both abstract constraints as well as the overall frequency of the binomial types. Results from comprehension tasks in Morgan and Levy (2016a) also showed that online processing patterns of highly frequent

binomials are directly shaped by their frequency.

One remaining question is why there should be a stronger regularization bias in the dative construction compared to binomials. Previous work has shown that the extent of regularization depends on cognitive load (Ferdinand et al., 2019). Learners tend to regularize more when the cognitive load of the specific learning tasks is high. Binomials and the dative construction have different levels of syntactic complexity. The ordering of a binomial expression only involves the two content words within the binomial. For instance, with *safe and sound*, a language user mainly needs to figure out whether to put *safe* or *sound* first. This is much easier than choosing the argument realization pattern for a dative construction, where a language user has to decide whether to use a V-NP-PP order, or a V-NP-NP order. Both orders have many more words and more nested hierarchical structures than binomials. Since the dative construction is structurally more complex, its ordering might involve more cognitive load than ordering the two content words in a binomial, which results in language users having more regularized preferences.

Further experiments on idiosyncrasy in other types of syntactic alternations such as adjective ordering or adverb placement, especially in a crosslinguistic context, would provide valuable insights into the existence and extent of regularization. Methodologically, the ILMs that I have adopted here assume that one learner only takes the output of one other learner as the input, whereas in reality language users learn from multiple sources at the same time. Smith et al. (2017) successfully approximated this learning process by letting the learner take input from more than one speaker within each generation. Future work should explore how different model types compare in their explanatory power with regard to regularization.

Chapter 8

Developing a Dependency Treebank for Hupa

8.1 The Hupa language and its documentation

While the availability of large-scale multilingual corpora has advanced quantitative studies of language structures, most of the corpora have been developed for already well-documented or extensively studied languages, in particular Indo-European languages, as are evident in the treebanks mainly used so far in this dissertation. This poses limitations on both the scope and validity of typological studies as well as on claimed language universals.

This chapter reports on efforts to develop a dependency treebank for the Hupa language, a Dene/Athabaskan language spoken in northwestern California. This was carried out by the project team led by Dr. Justin Spence, along with Dr. Kayla Palakurthy, Tyler Lee-Wynant and the author of this dissertation. Hupa is traditionally spoken in Hoopa Valley on the lower Trinity River in the current Humboldt County. Compared to other California Dene languages, Hupa is the most extensively documented one to date (Goddard, 1904; Sapir and Golla, 2001; Golla, 1984). The earliest traceable texts for Hupa can be found in Jeremiah Curtin's unpublished field notes from the late 1880s (now archived at the National Anthropological Archives). A large collection of audio recordings and transcriptions for Hupa from Mary Woodward, Victor Golla and Sean O'Neil is now archived at the Survey of California and Other Indian Languages at the University of

California, Berkeley (UC Berkeley).

The Hupa Language Documentation Project (HLDP) originally based at UC Berkeley has been devoted specifically to documentation work on Hupa over recent years. Since the currently available materials of Hupa have come from a variety of sources collected by different researchers over a century, the materials vary greatly in their quality of documentation. Therefore one of the central goals of the HLDP has been to assemble these diverse strands of Hupa documentation and compile them into one single corpus resource. Specifically, the corpus contains both written texts and transcriptions of spoken data. The written data include three published collections: Goddard (1904), Sapir and Golla (2001), and Golla (1984). The spoken data come from transcriptions of field recordings created with Hupa elder Mrs. Verdena Parker. She has produced a significant number of recordings of Hupa texts since early 2000, some of which was a result of collaborations with members of the Hoopa Valley Tribe or with linguists who have been working on HLDP. The corpus has been under development since 2008 and has now grown to over 36,000 glossed units, which include both single words and multi-word expressions. All glossed units have been normalized to a unified practical orthography.

Moreover, the corpus is fully concordanced with an associated lexical database. The lexical database was based initially on a learner-oriented print dictionary (Council, 1996). Along with the development of the Hupa corpus, the lexical database has been gradually expanded with new glossed units harvested from the corpus. It has also benefited significantly from consultations with Mrs. Parker who provided advice regarding detailed verb paradigms. Both the lexical database and the text corpus can be accessed through the Hupa Online Dictionary and Texts website. [1]

8.2 Annotation framework

One of the fundamental problems with developing syntactic annotations for less-studied languages like Hupa is that the syntax of these languages is still very much poorly understood. It is a general caveat with research on Dene languages that work on syntactic analysis has been lacking. Although documentation of Hupa has been carried out for

[1] `http://nalc.ucdavis.edu/hupa/hupa-lexicon.html`

decades, very few studies have tried to probe its constituent orderings (Spence, 2013, 2008; Newbold, 2010). These studies have focused on different samples of the documentation and have relied on manual analyses. Therefore one of the primary goals in developing the Hupa dependency treebank is to offer a more syntactically coded corpus to facilitate research on related topics in Hupa.

With that being said, designing syntactic annotations for a corpus presupposes that the correct syntactic analysis is already in sight. Unfortunately, this is not often the case with low-resource languages like Hupa. Therefore another crucial goal of the treebank development described here is to use annotation as a way to uncover syntactic relations and formalize analyses of Hupa syntax. As we encountered recurring phenomena from the corpus, we were able to gain a better understanding of the syntactic structures, and could refine our annotations iteratively in return.

Specifically, we adopted the dependency grammar framework for syntactic annotation. There are three main reasons why we used this specific grammatical framework. First, Hupa has relatively free word orders, which can be captured more elegantly by dependency representations than phrase structure rules (see Section 2.1).

Secondly, this framework allows us to take advantage of the annotation schema of UD, which can be tuned for language-specific features and variability. In particular, Dene languages in general are morphologically rich and there are still many remaining issues regarding the correct morphological segmentations and the functions of different morphemes (Holton and Lovick, 2008). The annotation guidelines of UD allow us to shy away from these issues for now, as the guidelines do not necessarily require the corpus to be morphologically parsed.

Thirdly, there has already been a dependency treebank under development for the Karuk language, another Native American language of northwestern California (Garrett et al., 2013, 2015). The Karuk text database also uses an XML format, the organization of which bears similarities to that of the Hupa database. This facilitates adapting the annotation guidelines already designed for Karuk to Hupa. In addition, although Karuk is not a Dene language, previous studies have discussed language contact effects between Karuk

and Hupa (Conathan, 2004). Thus learning from the annotation schema of the Karuk treebank would in the long run motivate a direct comparison of syntactic relationships in the two languages.

8.3 Implementation and workflow

To date, the project has completed annotations of 23 of the 74 texts in Sapir and Golla (2001), which is approximately 26% of the Spair collection and 13% of the full corpus overall.

Features	**Frequency**
glossed dictionary units	4700
sentences	814
tokens	6561
token types	2192

Table 8.1: Descriptive statistics of the annotated Hupa treebank.

Before dependency annotations were developed, the corpus was originally formatted following an XML schema, designed and carried out by Dr. Justin Spence and the author of this dissertation. The schema largely conform to the Text Encoding Initiative (TEI) [2] (Czaykowska-Higgins et al., 2014), which offers robust, flexible and widely-adopted representation standards for electronic text corpora. Our annotation procedures started by converting the XML representations of the corpus to the tab-delimited CoNLL-U format used by UD treebanks. This allows us to use the existing tools provided by UD such as Arborator, an online dependency grammar annotator (Gerdes, 2013). [3]

As an illustration, let us consider the following sentence in Hupa.

(1) na'te:dichwiw hay tsumehstł'o:n
she cried along the woman
'The woman cried as she went back along.'

[2] `https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf`
[3] `https://arborator.ilpga.fr`

In this sentence, *tsumehstł'o:n* ('woman') is the subject of the sentence, and it has the following XML representation in the original corpus. Its syntactic dependency relation with the verb is represented as *nsubj* within the $< ref >$ tag in the second line below, including a pointer to the verb's unique identifier in the database ($\#Sapir - 11 - 94$).

```
<w xml:id="Sapir-11-96">
 <ref type="dependentOf" target="#Sapir-11-94">nsubj</ref>
 <reg>
  <m>tsumehst\l{}'o:n</m>
 </reg>
</w>
```

After converting to the CoNLL-U format, the sentence would have a representation as shown in Figure 8.1. Each line represents the syntactic and documentation information for each token. Within each line, each non-empty column, from left to right, represents respectively for the token: its index in the sentence, its written form, its English translation, its part-of-speech tag, its id in the Hupa Online Dictionary and Texts website, the index of its syntactic head, its dependency relation with its syntactic head, and its unique identifier in the XML database.

1	na'te:dichwiw	She cried along	VERB	682	_	0	root	_	Sapir-11-94
2	hay	the	DET	3097	_	3	det	_	Sapir-11-95
3	tsumehstł'o:n	woman.	NOUN	3457	_	1	nsubj	_	Sapir-11-96

Figure 8.1: Example of dependency annotation in CoNLL-U format in the Hupa treebank.

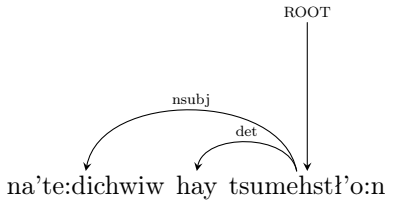
The annotation procedures were carried out by our project team in a rotating fashion. To ideally reach standard methods of annotation for common syntactic constructions, each text was annotated by two members of the team in order to ensure consistency. Once two annotators had completed a text, a third member compared the results, drew on comparisons with previously annotated sentences having similar structures, and reconciled discrepancies in the two annotation files. In certain cases we consulted specific UD corpora to determine how similar types of constructions are handled in the annotations of other languages. When there were conflicting annotations, after we made a decision,

we documented our decisions for future reference and shared them with the rest of the project team.

Annotation decisions were guided mainly by UD principles and dependency relations and relied heavily on the English translation of the Hupa data. Accordingly, our annotations also focused on content head, rather than function head. In the meantime, our decisions were refined iteratively as new structures were encountered in the corpus or in the ongoing recordings of stories produced by Mrs. Parker.

For each sentence, we first identified the head of the sentence, which typically is the main verb. We then proceeded to identify the dependency relation for each additional word in the sentence. This method works for relatively simple sentences like the one presented above. Its dependency representation would be as follows.

(2)

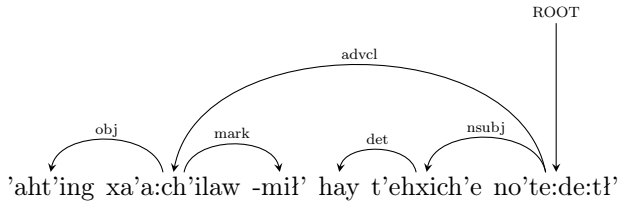


The annotation method also works equally well for more complex sentences, such as the following sentence from Sapir and Golla (2001). In this one, the head verb is *no'te:de:tł'* ('they all sat down'), which is the root of the sentence. It has a subject dependent *t'ehxich'e* ('girl') as well as an adverbial clause dependent. The head verb of the adverbial clause is *xa'a:ch'ilaw* ('he had done so'), which has both a subordinating enclitic dependent *-mił'* and a direct object *'aht'ing* ('all'). We used *mark* to represent the dependency relation between the head verb of the adverbial clause and the subordinating enclitic, as *mark* is typically applied for marking subordinate clauses in UD.

(3) 'aht'ing xa'a:ch'ilaw-mił' hay t'ehxich'e no'te:de:tł'

All (rocks) when he has done so (to them) the girls they all sat down

'When he had done this to all the rocks, every one of the girls sat down.'



Note that here the actual object of the head verb in the adverbial clause, *rocks* is not

expressed in the sentence. Following the principle of *promotion* by UD, when a content word is elided, an item that would routinely be analyzed as its dependent can take upon the dependency relation that the elided word would otherwise have had. Therefore here the determiner *'aht'ing* ('all'), which supposedly is the dependent of *rocks*, is promoted to be the object of *xa'a:ch'ilaw*.

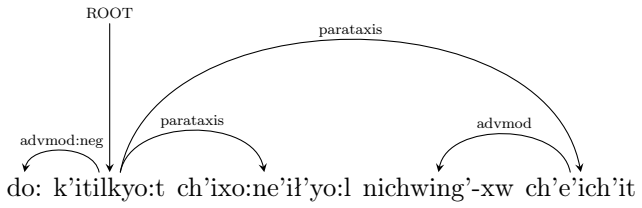
While identifications of dependency relations for cases like those mentioned above are comparatively simple and straightforward, there are many instances where the identifications are much more complex. We give three specific examples below in detail.

Although many sentences in the corpus have just one main verb, there are a number of cases that have additional verbs without any indication of coordinate or subordinate relations with the main verb. In such cases, we assigned the first verb of the sentence as the *root*. Subsequent verbs were analyzed as dependents of the root, with a dependency relation of *parataxis*. This dependency relation is ordinarily used for clauses that are ordered side by side to their head without bearing explicit syntactic coordination with or subordination to the head. For the following example in particular, since our annotations depend largely on the English translation, here the actual syntactic dependency relation between the clauses might be obscured by the semicolons in the translation, which indicates that *do: k'itilkyo:t* bears no syntactic relation with the clauses in the rest of the sentence. However, by treating such cases as *parataxis*, we were able to annotate them consistently, which makes it easier to locate them later on and modify the analysis if necessary. This example provides a good sense of one of our annotation principles, which is consistency, even though consistency might lead to annotation errors.

(4) do: k'itilkyo:t ch'ixo:ne'ił'yo:l nichwing'-xw ch'e'ich'it

One should not steal; they wish him bad luck; in a bad way he dies

'He doesn't steal; people swear at him, and he dies in a bad way.'

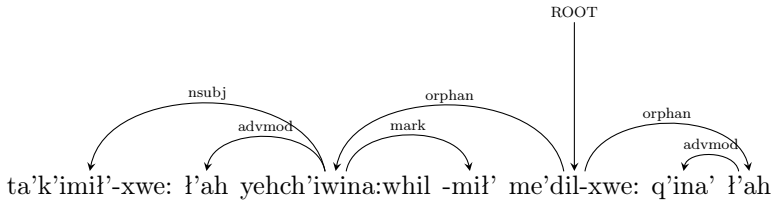


The second example of analytic difficulty involves ellipsis. As illustrated below, there is no main verb that means 'do' in the sentence to be the *root*, and a remnant nominal subject *me'dil-xwe:* is assigned as the *root* based on the principle of promotion, as discussed above. Other clausal level dependents of the elided main verb are now dependents of the subject *me'dil-xwe:* and their dependency relation with the head is *orphan*, which is used "in cases of head ellipsis where simple promotion would result in an unnatural and misleading dependency relation". [4] Here we give preference to the subject of the elided verb instead of the head verb of the adverbial clause in order to maintain structure of the main clause of the sentence.

(5) ta'k'imił'-xwe: ł'ah yehch'iwina:whil-mił' me'dil-xwe: q'ina' ł'ah

Ta'k'imił'ding people once every time he has come in Me'dilding people also once

'Each time the *ta'k'imił'xwe:* go in to dance once, the *me'dilxwe:* also do so once.'



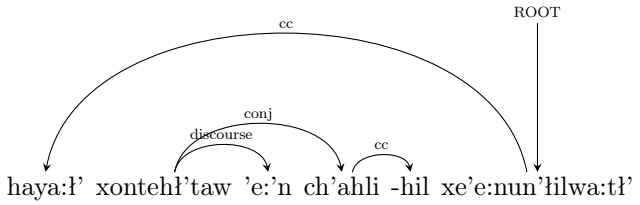
The last example illustrates a case in which we took shallow discourse information into consideration. The particle *'e:'n* ('for his part') is analyzed as the dependent of the preceding nominal *xontehł'taw* ('Coyote') with a *discourse* dependency relation for two reasons. First, *discourse* is intended for "interjections and other discourse particles and elements which are not clearly linked to the structure of the sentence, except in an expressive way". [5] Here *'e:'n* does typically express some degree of affective meaning, as it is sometimes glossed with an English interjection such as "Indeed!". Secondly, with this case particularly, the particle has a more well-defined discourse function, possibly indicating a shift in topic from the preceding context of this sentence. We did not treat the particle as the *discourse* dependent of the head verb of the sentence, given that its English translation suggests that semantically the particle is linked with *xontehł'taw* ('Coyote').

[4] https://universaldependencies.org/u/dep/orphan.html
[5] https://universaldependencies.org/u/dep/discourse.html

(6) haya:ł' xontehł'taw 'e:'n ch'ahli-hil xe'e:nun'łilwa:tł'

Then Coyote for his part together with Frog they had thrown each other away

'Coyote and Frog had walked out on each other.'



8.4 Applications to word order and future directions

This section presents some simple preliminary results of analyzing Hupa constituent orderings with the treebank. Though our annotations of the Hupa dependency treebank are still subject to revisions so far, we focused specifically on major constituent orders, as the annotations for these structures are unlikely to be substantially affected by subsequent changes to the designed annotation schema. The reason is that the identifications of clause-level grammatical relations such as subject and object are in general more straightforward and less error-prone in comparison to other types of grammatical relations.

Recall that it was mentioned in Chapter 1 that although Hupa is claimed to be predominantly SOV, the subject and the object can both appear after the head verb quite freely. Therefore in Hupa, (S)VO and (O)VS orders are not rare. As argued in Conathan (2004), whether the NP appears in preverbal or postverbal domains in Hupa depends on their discourse status. NPs that occur postverbally tend to be discourse-old (*given*) whereas NPs that occur preverbally are discourse-new. This observation goes against the Principle of Given-before-new attested in other studies that have mainly looked at Indo-European languages (Prat-Sala and Branigan, 2000; Arnold et al., 2000).

Later work by Newbold (2010) has confirmed the observation from Conathan (2004). Furthermore, Newbold (2010) quantified the relative distributions of preverbal vs. postverbal NPs in main clauses using a subset of the texts in Sapir and Golla (2001). The results showed that subject, direct object and oblique NPs occur in postverbal position approximately 28% of the time.

Although the Hupa treebank is not annotated with discourse information, we tried to replicate the results for NP distributions in main clauses from Newbold (2010) using our data, extended to adverbial clauses as well. The results show similar patterns in main clauses (Table 8.2). The patterns for NP distributions in adverbial clauses in addition (Table 8.3) correspond to those in Newbold (2010), in the sense that NPs predominantly appear before the head verb.

	Preverbal	**Postverbal**
subject	174 (69.6%)	76 (30.4%)
object	196 (70.3%)	83 (29.7%)
oblique	431 (77.7%)	124 (22.3%)
total	801 (73.9%)	283 (26.7%)

Table 8.2: Distribution of preverbal and postverbal NPs in main clauses.

	Preverbal	**Postverbal**
subject	45 (93.8%)	3 (6.3%)
object	29 (90.6%)	3 (9.4%)
oblique	63 (92.6%)	5 (7.4%)
total	137 (92.6%)	11 (7.4%)

Table 8.3: Distribution of preverbal and postverbal NPs in adverbial clauses.

After having seen the distributions of single NP, we looked at the orderings in clauses with exactly two NPs. Again, we focused on the relative order of the subject, direct object, and the oblique of the head verb. As shown in Table 8.4, when a clause has two NPs, they both prefer to appear preverbally, rather than postverbally.

Overall, though our development of the treebank and its refinements are still at a preliminary stage, the treebank so far has heuristic values for data exploration, making it a useful starting point for identifying distributional syntactic patterns. As we emphasized in the beginning, the current goal is not to arrive at the absolute correct annotations yet, especially since there are no gold-standard references, but instead to use annotation to

	NP_1 NP_2 **V**	NP_1 **V** NP_2	**V** NP_1 NP_2
{subject, object}	8 (36.3%)	13 (59.1%)	1 (4.5%)
{subject, oblique}	29 (34.5%)	50 (59.5%)	5 (6.0%)
{object, oblique}	47 (47.5%)	50 (50.5%)	2 (2.0%)
{oblique, oblique}	40 (55.6%)	30 (41.7%)	2 (2.8%)
Total	124 (44.8%)	143 (51.6%)	10 (3.6%)

Table 8.4: Distribution of two NPs in main clauses.

pave the way for a deeper understanding of Hupa syntax.

More generally, we hope to contribute to a broader conversation among Dene language scholars as well as other researchers who are also working on low-resource languages and are taking a corpus-based approach. Even though we might be applying different frameworks across the research community as of now, we hope to raise awareness that it would be ideal to share framework guidelines as well as annotation standards to talk about the possibility of adopting common annotation standards. These standards do not even have to use a dependency grammar framework. This would in turn help us to learn from each other and motivate historical-comparative research, which would not only be limited to syntax and structural reconstructions. For instance, the Wailaki language, another Dene language of California, is not documented to the same extent as Hupa. Yet it has drawn on comparative data from the Hupa corpus in the development of a language revitalization program (Begay et al., To appear).

What I have given above is the official, academic side of this chapter. Here finally I would like to point to some humanitarian aspects regarding research involving indigenous languages. Any language researchers, myself included, who need input from people who are native speakers of endangered and low-resource languages or who come from their community and tribes, should first respect their boundaries and their time. As researchers and hopefully scholars, we should always take care to recognize their efforts and acknowledge their contributions in both non-academic and academic contexts. It should always be the case that we remember to ask for permission before making data

recordings and sharing practices. When we are doing field recordings, we should respect the routines of the elders and be mindful about their needs (e.g. make sure they can take breaks, eat food and drink water). We should never pressure community members just in order to achieve personal academic goals, and should never lie or keep people from the language community outside of the loop when there are publications or grant application opportunities. I am grateful to Mrs. Verdena Parker for letting us work with her, for all her beautiful stories; to attendees at the 2018 Breath of Life held at UC Berkeley; and to Dr. Justin Spence, for including me in the project on Hupa. After all, science is more than just tables of numbers. Languages belong to the people.

Chapter 9

Conclusion: Looking Ahead

𝄞 *Yoda: Much to learn, you still have.*

Borrowing insights from language processing and language evolution, this dissertation has tried to address the broader question of why languages are the way they are by focusing on aspects of natural language syntax, situated within the dependency grammar framework. In particular, this project has aimed to model word ordering preferences both within and across languages using a data-driven approach. To do this, I have tested predictions of both abstract constraints and idiosyncratic biases motivated by long-standing theoretical principles as well as previous empirical findings. Overall, I have shown that while the effects of individual factors depend on the specific ordering patterns of different languages, the predictive power and direction of these constraints appear to be more dependent on whether the orderings are in the preverbal or the postverbal domains.

In more detail, I have examined the roles of dependency length, semantic closeness/argumenthood status, lexical frequency, contextual predictability, and word co-occurrence information at a large crosslinguistic scale (see Chapter 3, 4 and 5). The results indicated that while the first two factors are significant typological determinants of ordering preferences, the latter three frequency-based measures are much less effective. Between dependency length and semantic closeness, my findings have shown that the preference for shorter dependencies is much stronger in postverbal orderings compared to preverbal

orderings, a contrast which is less constrained by particular language types. On the other hand, even though in general the effect of semantic closeness is weaker than dependency length, its role is much more consistent across languages. In certain preverbal domains where dependency length has a very weak or no effect, semantic closeness appears to be a more robust predictor.

Additionally, I have also explored word order flexibility (see Chapter 6), an overlooked dimension in empirical studies of syntactic typology, as a potential factor to explain the crosslinguistic variation of DLM. I have presented evidence that across languages, syntactic constructions with longer dependencies have more variable orderings, and that there is more flexibility in constituent orders before than after the head verb.

Besides the variety of abstract factors that I have studied, I have also investigated idiosyncrasy biases in syntactic orderings. Specifically, I have looked at verb bias in the dative construction in English (see Chapter 7). The results showed that the higher the overall frequency of the syntactic construction, the stronger are the preferences that language users will have for one syntactic alternation over the other. Furthermore, I have demonstrated that this pattern arises from the interactions between continuous cultural transmission and language processing, adopting methodologies from the literature on language evolution.

Finally, I adopted dependency syntax representations to formalize and model word orders in Hupa (see Chapter 8), and called for unified efforts, attention and care when studying indigenous languages. In this last chapter, I would like to talk about caveats in both the dissertation as well as the current research on syntactic typology in the field, and discuss possible future directions.

9.1 Corpus observations and computation vs. human preferences

While the modeling in this dissertation predicts the orderings observed in the corpora, it is valuable to keep in mind that corpus observations are not necessarily the orders preferred by language users. When none of the factors under examinations are effective,

it is possible that the observed ordering and its structural variant are equally preferred, or that there is only a small preference for one syntactic alternative over another.

To reveal the correspondences as well as mismatches between corpus observations and human preferences more generally, it would be ideal to corroborate findings here with acceptability judgement tasks on a crosslinguistic scale. This methodology has been applied previously in different studies (Bresnan and Ford, 2010; Bresnan, 2007) to probe whether language users' syntactic knowledge is probabilistic, though mostly in English with some exceptions (Divjak et al., 2016). These experiments constructed statistical models based on sentences extracted from corpora and showed that predictions of the model correlate well with acceptability ratings of the same sentences.

One question that my advisor has asked me during my qualifying exam is that since all of the quantitative measures are methods of number counting (like those in Chapter 5), would they truly capture what is going on in language use? To be honest, I cannot say that the answer is a solid 'yes'. These computational methods are at best approximations for testing the predictions of traditional linguistic theories. Future research should shift more attention to online production studies to test the factors outlined in this dissertation. For example, given the much weaker effect for DLM in preverbal domains presented in Chapter 3, it would be interesting to test whether that is also the case when speakers are trying to order the constituents in their utterances in real time. With everything else being equal (ideally), one could also compare the predictive power of the Principle of Argument Closer versus the Principle of Argument Precedence in different language types with psycholinguistic methods. These directions require carefully designed stimuli, which can potentially be adapted from corpus observations.

The question of whether ordering instances found in corpora or computation based on corpus statistics would match human preferences also extends to other linguistic notions, such as (non-)canonical orders and flexibility. As mentioned in Namboodiripad (2019), the canonical ordering in a language is not always its dominant word order. Instead, what determines an order as the canonical one depends on the discourse context of the structure, not just on its corpus frequency. Similarly with flexibility, how flexible the

word orders of a language are should not solely depend on how often these orders are observed in corpora, as is done in Futrell et al. (2015b), Levshina (2019) and Chapter 6 of this dissertation. One alternative way, as proposed in Namboodiripad (2019), is to use an offline acceptability judgement task to measure the flexibility in constituent orderings. In this approach, if a sentence is rated as more accessible, it is considered more flexible.

9.2 Gradient syntactic ordering preferences

The discussions in Section 9.1 align with a broader point made by previous studies (Bresnan and Ford, 2010; Bresnan, 2007; Bresnan et al., 2007; Morgan and Levy, 2016a) that word order preferences are gradient. Imagine a syntactic construction with two alternative sturcutures, A and B. Instead of making a binary choice between whether A or B is more preferred, as is the approach taken in this dissertation, future work can address how much more likely A is preferred to B or vice versa. This would offer greater insights into the upper bound or lower bound of the magnitude effect for different constraints. For instance, as quantified in Hawkins (1994) and Wiechmann and Lohmann (2013) as well as in Chapter 3, in postverbal double PP orderings in English, as the length difference between the two PPs increases, the effect of dependency length on ordering choices grows stronger accordingly. For example, the preference for DLM is in general much weaker when the lengths of the two PPs differ by only 1 to 3 tokens, in contrast to when the length difference is much larger.

Following this line of work, a natural extension of Chapter 4 or Chapter 5 could be, instead of predicting which order is more preferred, to construct a model that predicts the overall dependency distance difference between the observed ordering and its structural variant, given the relative dependency length between the head verb and each PP, along with other constraints that have been quantified. This would offer more direct evidence for the gradience in the ordering preferences.

9.3 Individual structures vs. language as a whole

While this dissertation has tried to address why natural language syntax is the way it is via investigations of individual structures, recently researchers have developed another

approach that compares the observed grammar of a language as a whole to that of randomly generated baselines (Gildea and Temperley, 2010; Temperley, 2007; Futrell et al., 2015a; Hahn et al., 2020). These authors have argued that grammars of human languages are optimized to achieve efficient communication.

But what is communicative efficiency? Could one ever come up with estimates that reasonably approximate it? I believe that as language users, we hope to be rational and to efficiently communicate with each other, but a concept so broad is unlikely to be largely reflected by quantitative methods from information theory. Additionally, aggregated measures of distributional patterns in a language as a whole miss more fine-grained information of syntactic constructions at different levels of granularity, given that the realizations of these constructions are driven by multifactorial constraints.

Regarding individual structures, while this dissertation has reviewed and tested a series of factors, it is clear that not all of them can be generalized to other syntactic constructions beyond those considered. As an illustration, consider adjective orderings. The case that has been studied the most is double prenominal adjective orderings in English, with some crosslinguistic extensions (Dyer, 2017; Samonte and Scontras, 2019; Sproat and Shih, 1991; Gulordava et al., 2015; Gulordava and Merlo, 2015b). The traditional literature has tried to explain adjective orderings with various proposals, including semantic (Dixon, 1982) and syntactic (Cinque, 1994, 2010, 2014) hierarchies as well as cognitive factors including apparentness (Sproat and Shih, 1991), frequency (Wulff and Gries, 2015), word co-occurrence information (Hahn et al., 2018; Futrell, 2019; Martin, 1969), subjectivity (Scontras et al., 2017, 2019), integration cost (Dyer, 2017) and information gain (Futrell et al., 2020a). While some of these measures have been shown to achieve decent prediction accuracy, how these measures interact across languages remains to be investigated.

9.4 (First) Language learning

While most work on language typology has focused on social and cognitive factors from both diachronic and synchronic perspectives, one area that has not received the atten-

tion that it deserves is language learning and how this affects structural variation across languages. Studies of the patterns and trajectories whereby children learn and reproduce structural alternatives from a crosslinguistic angle could potentially shed light on how linguistic variations turn out to be the way they are in the first place.

For instance, given that the preference for shorter dependencies is motivated by cognitive load, and that children have lower working memory load in comparison to adults, one could look at the trajectory of DLM in first language acquisition, possibly starting with corpora from the CHILDES database (MacWhinney, 2000). There might be observations of greater extent of DLM in early utterances of child speech, with this extent possibly decreasing over time and correlating negatively with the age of the child.

It is also worth attempting to formalize DLM as a measure for levels of language development. For example, future studies could investigate whether systematic differences in DLM exist in speech produced by children with typical development compared to that produced by children with language impairment (Prud'hommeaux et al., 2014). One could potentially compare dependency length with other standard syntactic features that have been adopted in automatic measurements of language development (Lubetich and Sagae, 2014; Sahakian and Snyder, 2012).

9.5 Prediction in English-like languages is not everything

The field of language science has gradually shifted to being more open to quantitative methods and computational modeling, ranging from simple regression models to neural language models. For instance, there has been a plethora of work on structure prediction with neural networks, asking whether neural models have captured phenomena X (Linzen, 2019; Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2019; Futrell and Levy, 2019; Futrell et al., 2019). Yet most of these studies are English-centric.

While English is admittedly interesting, it is not more *special* than any other languages. As pointed out in previous chapters, we should be cautious when drawing conclusions based on the data type and sample size in the experiments. Having tables/figures of

numbers does not necessarily make the study more scientific, and making claims about human language based on English alone is certainly not adequate. While most of the languages in this dissertation behave like English, I hope that my findings with the mixed-types as well as head-final languages will open up new venues for future research. I hope to be able to account for the patterns that my experiments have not yet been able to explain, and to discover new structural features of these languages that have not been documented before. These are some of the research goals that I have set for myself and for the field as a whole.

Appendix A

A.1 Descriptive statistics of total VP instances in Chapter 3

The number of total VP instances examined in each language under different case scenarios is presented below.

Language	Total VP instances	Language family and genus
Danish	198	Indo-European, Germanic
Norwegian	1323	Indo-European, Germanic
Swedish	488	Indo-European, Germanic
Arabic	561	Afro-Asiatic, Semitic
Hebrew	964	Afro-Asiatic, Semitic
Greek	127	Indo-European, Greek
Indonesian	357	Austronesian, Malayo-Sumbawan
Galician	221	Indo-European, Romance
Latvian	59	Indo-European, Baltic
Irish	83	Indo-European, Celtic
Serbian	177	Indo-European, Slavic
Slovak	114	Indo-European, Slavic

Table A.1: Descriptive statistics of total VP instances for languages with head-initial PPs after the head verb.

Language	Total VP instances	Language family and genus
Afrikaans	92	Indo-European, Germanic
Persian	870	Indo-European, Iranian
Chinese	111	Sino-Tibetan

Table A.2: Descriptive statistics of total VP instances for Afrikaans, Persian and Chinese.

Language	Total VP instances	Language family and genus
Japanese	166	Japanese
Hindi	1152	Indo-European, Indic
Urdu	294	Indo-European, Indic

Table A.3: Descriptive statistics of total VP instances for languages with head-final PPs before the verb.

Language	Total VP instances	Language family and genus
English	1250	Indo-European, Germanic
German	6400	Indo-European, Germanic
Dutch	459	Indo-European, Germanic
Bulgarian	164	Indo-European, Slavic
Ukrainian	237	Indo-European, Slavic
Slovenian	200	Indo-European, Slavic
Russian	2179	Indo-European, Slavic
Czech	3093	Indo-European, Slavic
Croatian	305	Indo-European, Slavic
Polish	936	Indo-European, Slavic
French	2236	Indo-European, Romance
Spanish	2672	Indo-European, Romance
Portuguese	504	Indo-European, Romance
Romanian	564	Indo-European, Romance
Italian	2188	Indo-European, Romance
Catalan	714	Indo-European, Romance

Table A.4: Descriptive statistics of total VP instances when head-initial PPs appear after the verb, in languages where head-initial PPs can also appear before the verb.

Language	Total VP instances	Language family and genus
English	55	Indo-European, Germanic
German	14892	Indo-European, Germanic
Dutch	625	Indo-European, Germanic
Bulgarian	77	Indo-European, Slavic
Ukrainian	121	Indo-European, Slavic
Slovenian	248	Indo-European, Slavic
Russian	1655	Indo-European, Slavic
Czech	1640	Indo-European, Slavic
Croatian	151	Indo-European, Slavic
Polish	318	Indo-European, Slavic
French	151	Indo-European, Romance
Spanish	182	Indo-European, Romance
Portuguese	70	Indo-European, Romance
Romanian	62	Indo-European, Romance
Italian	330	Indo-European, Romance
Catalan	115	Indo-European, Romance

Table A.5: Descriptive statistics of total VP instances when head-initial PPs appear before the verb, in languages where head-initial PPs can also appear after the verb.

A.2 Descriptive statistics of total VP instances in Chapter 6

The number of total VP instances examined in each language under different case scenarios is presented below.

Language	Total VP instances	Language family and genus
Arabic	1060	Afro-Asiatic, Semitic
Hebrew	1108	Afro-Asiatic, Semitic
Indonesian	492	Austronesian, Malayo-Sumbawan
Greek	290	Indo-European, Greek
Wolof	158	Niger-Congo, Northern Atlantic
Irish	112	Indo-European, Celtic
Latvian	217	Indo-European, Baltic
Danish	717	Indo-European, Germanic
English	3153	Indo-European, Germanic
Swedish	1279	Indo-European, Germanic
Norwegian	4003	Indo-European, Germanic
Portuguese	1242	Indo-European, Romance
Romanian	946	Indo-European, Romance
French	4011	Indo-European, Romance
Galician	550	Indo-European, Romance
Italian	4304	Indo-European, Romance
Bulgarian	630	Indo-European, Slavic
Serbian	393	Indo-European, Slavic
Ukrainian	634	Indo-European, Slavic
Finnish	185	Uralic, Finnic

Table A.6: Descriptive statistics of total VP instances for languages with NP and PP after the head verb.

Language	Total VP instances	Language family and genus
Afrikaans	225	Indo-European, Germanic
Persian	953	Indo-European, Iranian
Japanese	648	Japanese
Hindi	1896	Indo-European, Indic
Urdu	670	Indo-European, Indic

Table A.7: Descriptive statistics of total VP instances for languages with NP and PP before the head verb.

Language	Total VP instances	Language family and genus
Catalan	2324	Indo-European, Romance
Spanish	5125	Indo-European, Romance
Dutch	601	Indo-European, Germanic
German	9837	Indo-European, Germanic
Croatian	694	Indo-European, Slavic
Czech	7084	Indo-European, Slavic
Polish	2050	Indo-European, Slavic
Russian	4743	Indo-European, Slavic
Slovak	401	Indo-European, Slavic
Slovenian	341	Indo-European, Slavic
Estonian	489	Uralic, Finnic

Table A.8: Descriptive statistics of total VP instances when the NP and PP appear after the head verb, in languages where they can also appear before the head verb.

Language	Total VP instances	Language family and genus
Catalan	602	Indo-European, Romance
Spanish	503	Indo-European, Romance
Dutch	562	Indo-European, Germanic
German	14612	Indo-European, Germanic
Croatian	114	Indo-European, Slavic
Czech	1650	Indo-European, Slavic
Polish	190	Indo-European, Slavic
Russian	541	Indo-European, Slavic
Slovak	132	Indo-European, Slavic
Slovenian	136	Indo-European, Slavic
Estonian	179	Uralic, Finnic

Table A.9: Descriptive statistics of total VP instances when the NP and PP appear before the head verb, in languages where they can also appear after the head verb.

A.3 Detailed results for ordering flexibility and DLM in Chapter 6

Language	**Ordering flexibility**	**DLM**
Arabic	0.79 (0.75, 0.83)	1.56 (1.49, 1.63)
Hebrew	0.94 (0.92, 0.96)	1.22 (1.17, 1.28)
Indonesian	0.09 (0.04, 0.15)	1.27 (1.09, 1.47)
Greek	0.82 (0.74, 0.89)	1.74 (1.63, 1.88)
Wolof	0.45 (0.29, 0.60)	1.62 (1.45, 1.82)
Irish	0.67 (0.49, 0.81)	0.78 (0.60, 0.93)
Latvian	0.60 (0.47, 0.72)	1.27 (1.11, 1.45)
Danish	0.32 (0.25, 0.39)	2.05 (1.95, 2.21)
English	0.17 (0.14, 0.20)	1.66 (1.58, 1.78)
Swedish	0.26 (0.21, 0.31)	1.21 (1.11, 1.33)
Norwegian	0.38 (0.35, 0.41)	1.38 (1.33, 1.44)
Portuguese	0.84 (0.80, 0.87)	2.00 (1.93, 2.07)
Romanian	0.80 (0.75, 0.84)	1.53 (1.47, 1.60)
French	0.81 (0.79, 0.83)	1.60 (1.56, 1.63)
Galician	0.86 (0.80, 0.90)	1.73 (1.65, 1.82)
Italian	0.85 (0.84, 0.87)	1.76 (1.73, 1.79)
Bulgarian	0.57 (0.49, 0.64)	1.28 (1.20, 1.38)
Serbian	0.39 (0.29, 0.48)	1.54 (1.43, 1.69)
Ukrainian	0.91 (0.87, 0.95)	1.53 (1.45, 1.62)
Finnish	0.85 (0.75, 0.93)	1.16 (1.03, 1.33)
Average	**0.62**	**1.49**

Table A.10: Ordering flexibility and DLM in languages with postverbal NP and PP, measured with entropy.

Language	Ordering flexibility	DLM
Afrikaans	0.88 (0.80, 0.95)	0.63 (0.52, 0.77)
Persian	0.98 (0.96, 0.99)	0.40 (0.35, 0.46)
Japanese	0.77 (0.71, 0.83)	1.05 (0.97, 1.13)
Hindi	0.87 (0.84, 0.89)	0.58 (0.54, 0.61)
Urdu	0.85 (0.80, 0.90)	0.67 (0.61, 0.74)
Average	**0.87**	**0.67**

Table A.11: Ordering flexibility and DLM in languages with preverbal NP and PP, measured with entropy.

Language	Postverbal flexibility	Preverbal flexibility	Postverbal DLM	Preverbal DLM
Catalan	0.78 (0.75, 0.81)	0.23 (0.15, 0.29)	1.54 (1.49, 1.58)	1.34 (1.06, 1.62)
Spanish	0.76 (0.74, 0.78)	0.42 (0.33, 0.50)	1.71 (1.68, 1.75)	1.35 (1.14, 1.56)
Dutch	0.94 (0.90, 0.97)	1.00 (0.99, 1.00)	0.97 (0.90, 1.05)	0.15 (0.08, 0.23)
German	1.00 (1.00, 1.00)	0.99 (0.99, 0.99)	0.78 (0.76, 0.80)	-0.21 (-0.23, -0.20)
Croatian	0.42 (0.35, 0.49)	0.98 (0.93, 1.00)	1.53 (1.43, 1.64)	0.63 (0.38, 0.94)
Czech	1.00 (1.00, 1.00)	0.99 (0.98, 1.00)	1.30 (1.28, 1.33)	-0.14 (-0.20, -0.08)
Polish	0.88 (0.85, 0.90)	0.98 (0.94, 1.00)	1.64 (1.58, 1.69)	-0.21 (-0.36, -0.06)
Russian	0.90 (0.89, 0.92)	0.95 (0.92, 0.98)	1.55 (1.5,2 1.58)	0.30 (0.21, 0.39)
Slovak	0.96 (0.93, 0.99)	0.98 (0.93, 1.00)	1.32 (1.21, 1.46)	0.13 (-0.18, 0.41)
Slovenian	0.80 (0.73, 0.87)	0.95 (0.87, 0.99)	1.14 (1.04, 1.25)	0.08 (-0.11, 0.28)
Estonian	0.97 (0.95, 0.99)	1.00 (0.98, 1.00)	0.79 (0.71, 0.87)	0.13 (-0.03, 0.31)
Average	**0.86**	**0.85**	**1.30**	**0.32**

Table A.12: Ordering flexibility and DLM in languages with both postverbal and preverbal NP and PP, measured with entropy.

References

S. P. Abney. *The English Noun Phrase in its Sentential Aspect*. PhD thesis, Massachusetts Institute of Technology, 1987.

J. E. Arnold, A. Losongco, T. Wasow, and R. Ginstrom. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55, 2000.

I. Arnon. What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, 42(2):274–277, 2015.

I. Arnon and N. Snider. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67 – 82, 2010.

E. L. Bavin. Locative terms and Warlpiri acquisition. *Journal of Child Language*, 17(1): 43–66, 1990.

K. Begay, J. Spence, and T. Cheryl. Teaching Wailaki: Archives, Interpretation, and Collaboration. In A. Link, P. Spero, and A. Shelton, editors, *Translating Across Time and Space*. Lincoln: University of Nebraska Press, To appear.

O. Behaghel. *Deutsche Syntax: eine geschichtliche Darstellung*. Heidelberg: Carl Winters Universitätsbuchhandlung, 1932.

E. M. Bender. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, 2009. Association for Computational Linguistics.

S. B. Benor and R. Levy. The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language*, 82(2):233–278, 2006.

E. Berlage. *Noun Phrase Complexity in English*. Cambridge University Press, 2014.

A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G. Kim, M. A. Marcinkiewicz, and B. Schasberger. Bracketing guidelines for Treebank II style Penn Treebank project. *University of Pennsylvania*, 97:100, 1995.

D. Blasi, R. Cotterell, L. Wolf-Sonkin, S. Stoll, B. Bickel, and M. Baroni. On the Distribution of Deep Clausal Embeddings: A Large Cross-linguistic Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943, Florence, Italy, July 2019. Association for Computational Linguistics.

J. K. Bock. Syntactic persistence in language production. *Cognitive Psychology*, 18(3): 355 – 387, 1986.

J. K. Bock and D. E. Irwin. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, 19(4):467–484, 1980.

J. K. Bock and R. K. Warren. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47 – 67, 1985.

J. Bresnan. *Lexical-Functional Syntax*. Blackwell textbooks in linguistics, 2001.

J. Bresnan. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in Search of its Evidential Base*, 96:77–96, 2007.

J. Bresnan and M. Ford. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1):168–213, 2010.

J. Bresnan, A. Cueni, T. Nikitina, and R. H. Baayen. Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94. KNAW, 2007.

P.-C. Bürkner et al. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

J. Bybee. From Usage to Grammar: The Mind's Response to Repetition. *Language*, 82 (4):711–733, 2006.

J. Bybee and P. Hopper. Introduction to frequency and the emergence of linguistic structure. In J. Bybee and P. Hopper, editors, *Frequency and the emergence of linguistic structure*, pages 1–24. Amsterdam: John Benjamins, 2001.

J. L. Bybee. *Morphology: A Study of the Relation between Meaning and Form*, volume 9. John Benjamins Publishing, 1985.

L. Cai. The Semantic Functions of Prepositions and Postpositions in Chinese Spatial Circumpositions – A Perspective from Language Typology. In P. Liu and Q. Su, editors, *Chinese Lexical Semantics*, pages 248–257, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

T. Cameron-Faulkner, E. Lieven, and M. Tomasello. A construction based analysis of child directed speech. *Cognitive Science*, 27(6):843–873, 2003.

F. Chang. Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61(3): 374–397, 2009.

H.-W. Choi. Length and Order: A Corpus Study of Korean Dative-Accusative Construction. *Discourse and Cognition*, 14(3):207–227, 2007.

N. Chomsky. *Language and the Mind*. Harcourt, Brace, & World, 1968.

N. Chomsky. *The Logical Structure of Linguistic Theory*. Springer, 1975.

N. Chomsky. *Lectures on Government and Binding: The Pisa Lectures*. Number 9. Walter de Gruyter, 1993.

N. Chomsky. *The Minimalist Program*. Cambridge, Massachusetts: MIT Press, 1995.

M. H. Christiansen and N. Chater. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509, 2008.

M. H. Christiansen and S. Kirby. Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300–307, 2003.

K. Christianson and F. Ferreira. Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, 98(2):105–135, 2005.

G. Cinque. On the evidence for partial N-movement in the Romance DP. In *Paths towards universal grammar. Studies in honor of Richard S. Kayne*. Georgetown University Press, 1994.

G. Cinque. *The Syntax of Adjectives: A Comparative Study*, volume 57. Cambridge, Massachusetts: MIT Press, 2010.

G. Cinque. The semantic classification of adjectives. A view from syntax. *Studies in Chinese Linguistics, 35.(1) 3-32*, 2014.

C. J. Clifton, L. Frazier, and C. Connine. Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 23(6):696–708, 1984.

L. Conathan. *The Linguistic Ecology of Northwestern California: Contact, Functional Convergence, and Dialectology*. Doctoral dissertation, University of California, Berkeley, 2004.

R. Cotterell and J. Eisner. A Deep Generative Model of Vowel Formant Typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 37–46, New Orleans, June 2018.

R. Cotterell, C. Kirov, M. Hulden, and J. Eisner. On the Complexity and Typology of Inflectional Morphological Systems. *Transactions of the Association for Computational Linguistics*, 7:327–342, Mar. 2019.

H. V. T. Council. Hupa Language Dictionary. *compiled by Victor Golla*, 1996.

J. Culbertson and P. Smolensky. A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, 36(8):1468–1498, 2012.

J. Culbertson, P. Smolensky, and G. Legendre. Learning biases predict a word order universal. *Cognition*, 122(3):306–329, 2012.

P. W. Culicover and R. Jackendoff. *Simpler Syntax*. Oxford University Press, Oxford, 2005.

E. Czaykowska-Higgins, M. D. Holmes, and S. M. Kell. Using TEI for an endangered language lexical resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation & Conservation*, 8:1–37, 2014.

M.-C. de Marneffe and J. Nivre. Dependency Grammar. *Annual Review of Linguistics*, 5 (1):197–218, 2019.

K. De Smedt. Parallelism in Incremental Sentence Generation. *Parallel Natural Language Processing*, pages 421–447, 1994.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

D. Divjak, E. Dąbrowska, and A. Arppe. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics*, 27(1):1–33, 2016.

R. M. Dixon. *Where have All the Adjectives Gone?* De Gruyter Mouton, Berlin, Boston, 1982.

M. S. Dryer. SVO languages and the OV: VO typology. *Journal of Linguistics*, 27(2): 443–482, 1991.

M. S. Dryer. Word order. *Language typology and syntactic description*, 1:61–131, 2007.

M. S. Dryer. Order of Subject, Object and Verb. In M. S. Dryer and M. Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/81`.

M. S. Dryer and M. Haspelmath, editors. *The World Atlas of Language Structures Online.* Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/`.

W. E. Dyer. *Minimizing Integration Cost: A General Theory of Constituent Order*. PhD thesis, University of California, Davis, 2017.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

S. F. Ehrlich and K. Rayner. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655, 1981.

N. C. Ellis. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188, 2002.

N. J. Enfield. 'Case relations' in Lao, a radically isolating language. In *The Oxford Handbook of Case*, pages 808–819. Oxford University Press, 2009.

P. Faghiri and P. Samvelian. Word order preferences and the effect of phrasal length in SOV languages: evidence from sentence production in Persian. *Glossa: a journal of general linguistics*, 2020.

G. Fenk-Oczlon. Word frequency and word order in freezes. *Linguistics*, 27:517–556, 1989.

V. Ferdinand, S. Kirby, and K. Smith. The cognitive roots of regularization in language. *Cognition*, 184:53–68, 2019.

F. Ferreira. Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, 33(6):715–736, 1994.

F. Ferreira and M. W. Lowder. Prediction, information structure, and good-enough language processing. *Psychology of Learning and Motivation*, 65:217–247, 2016.

F. Ferreira and B. Swets. How Incremental is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, 46(1):57 – 84, 2002.

V. S. Ferreira and H. Yoshita. Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research*, 32(6):669–692, 2003.

R. Futrell. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France, Aug. 2019. Association for Computational Linguistics.

R. Futrell and R. Levy. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

R. Futrell and R. P. Levy. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 50–59, 2019.

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

R. Futrell, K. Mahowald, and E. Gibson. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden, Aug. 2015b. Uppsala University, Uppsala, Sweden.

R. Futrell, R. Levy, and E. Gibson. Generalizing dependency distance: Comment on

"Dependency distance: A new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews*, 21:197–199, 2017.

R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

R. Futrell, W. Dyer, and G. Scontras. What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2003–2012, Online, July 2020a. Association for Computational Linguistics.

R. Futrell, E. Gibson, and R. P. Levy. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814, 2020b.

S. Gahl and S. M. Garnsey. Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language*, 80:748 – 775, 2004.

N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16):E3635–E3644, 2018.

S. M. Garnsey, N. J. Pearlmutter, E. Myers, and M. A. Lotocky. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93, 1997.

A. Garrett, C. Sandy, E. Maier, L. Mikkelsen, and P. Davidson. Developing the Karuk Treebank. In *Fieldwork Forum, Department of Linguistics, UC Berkeley*, 2013.

A. Garrett, E. Maier, L. Mikkelsen, and C. Sandy. Exploring Karuk morphology in a parsed text corpus, 2015.

K. Gerdes. Collaborative Dependency Annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97, Prague, Czech Republic, Aug. 2013. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

J. Ghosh, Y. Li, R. Mitra, et al. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383, 2018.

E. Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1 – 76, 1998.

E. Gibson. The Dependency Locality Theory: A Distance-based Theory of Linguistic Complexity. In A. Marantz, Y. Miyashita, and W. O'Neil, editors, *Image, language, brain*, pages 95–126. Cambridge, Massachusetts: MIT Press, 2000.

E. Gibson and H.-H. I. Wu. Processing Chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2):125–155, 2013.

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

E. Gibson, R. Futrell, S. Piantadosi, I. Dautriche, K. Mahowald, L. Bergen, and R. Levy. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019.

D. Gil. Chapter 19 - Isolating-Monocategorial-Associational Language. In H. Cohen and C. Lefebvre, editors, *Handbook of Categorization in Cognitive Science (Second Edition)*, pages 471 – 510. Elsevier, San Diego, second edition edition, 2017.

D. Gildea and T. F. Jaeger. Human languages order information efficiently. *arXiv preprint arXiv:1510.02823*, 2015.

D. Gildea and D. Temperley. Optimizing Grammars for Minimum Dependency Length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*,

pages 184–191, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

D. Gildea and D. Temperley. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310, 2010.

F. Ginter, J. Hajič, J. Luotolahti, M. Straka, and D. Zeman. CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings, 2017. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

P. E. Goddard. Hupa texts. In *University of California Publications in American Archaeology and Ethnology*, number 2, pages 89–368. 1904.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP*, pages 517–520. IEEE, 1992.

A. E. Goldberg. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224, 2003.

A. E. Goldberg. The nature of generalization in language. *Cognitive Linguistics*, 20(1): 93–127, 2009.

A. E. Goldberg. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press, 2019.

V. Golla. *Hupa Stories, Anecdotes, and Conversations*. Hoopa, CA: Hoopa Valley Tribe, 1984.

J. H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg, editor, *Universals of Language*, pages 73–113. Cambridge, Massachusetts: MIT Press, 1963.

J. H. Greenberg. *Universals of Language*. Cambridge, Massachusetts: MIT Press, 1966.

S. T. Gries. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, 34(4):365–399, 2005.

S. T. Gries and N. C. Ellis. Statistical measures for usage-based linguistics. *Language Learning*, 65(S1):228–255, 2015.

K. Gulordava. *Word order variation and dependency length minimisation: a cross-linguistic computational approach.* PhD thesis, University of Geneva, 2018.

K. Gulordava and P. Merlo. Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden, Aug. 2015a. Uppsala University, Uppsala, Sweden.

K. Gulordava and P. Merlo. Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, 2015b.

K. Gulordava, P. Merlo, and B. Crabbé. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 477–482, Beijing, China, July 2015. Association for Computational Linguistics.

J. K. Gundel. Universals of topic-comment structure. *Studies in Syntactic Typology*, 17 (1):209–239, 1988.

M. Gustafsson. The frequency and 'frozenness' of some English binomials. *Neuphilologische mitteilungen*, 77:623–637, 1976.

M. Hahn, J. Degen, N. Goodman, D. Jurafsky, and R. Futrell. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

M. Hahn, D. Jurafsky, and R. Futrell. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353, 2020.

J. Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

K. Hale. Preliminary remarks on configurationality. In *Proceedings of the North Eastern Linguistic Society*, volume 12, pages 86–96, 1982.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

S. E. Haviland and H. H. Clark. What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13(5):512 – 521, 1974.

J. A. Hawkins. *Word Order Universals*. New York: Academic Press, 1983.

J. A. Hawkins. *A Comparative Typology of English and German: Unifying the contrasts*. London: Croom Helm, and Austin: University of Texas Press, 1986.

J. A. Hawkins. A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, 21(2): 223–261, 1990.

J. A. Hawkins. *A Performance Theory of Order and Constituency*, volume 73. Cambridge University Press, Cambridge, 1994.

J. A. Hawkins. The relative order of prepositional phrases in English: Going beyond Manner–Place–Time. *Language Variation and Change*, 11(3):231–266, 1999.

J. A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford, 2004.

J. A. Hawkins. Processing typology and why psychologists need to know about it. *New Ideas in Psychology*, 25(2):87–107, 2007.

J. A. Hawkins. *Cross-linguistic Variation and Efficiency*. Oxford University Press, Oxford, 2014.

J. A. Hawkins. Typological Variation and Efficient Processing. In *The Handbook of Language Emergence*, chapter 10, pages 215–236. John Wiley Sons, Ltd, 2015.

J. A. Hawkins. Word-external properties in a typology of Modern English: A comparison with German. *English Language & Linguistics*, 23(3):701–727, 2019.

B. Heinzerling and M. Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

G. Holton and O. Lovick. Evidentiality in Dena'ina Athabaskan. *Anthropological linguistics*, 50(3/4):292–323, 2008.

M. R. Hovav and B. Levin. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, 44(1):129–167, 2008.

R. Hudson. Measuring syntactic difficulty. *Draft of manuscript*, 1995. URL `http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf`.

C. L. Hudson Kam and E. L. Newport. Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, 1(2):151–195, 2005.

R. Jackendoff. *X-bar Syntax: A Study of Phrase Structure*. Linguistic inquiry monographs 2. Cambridge, Massachusetts: MIT Press, 1977.

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

T. F. Jaeger and R. P. Levy. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856, 2007.

T. F. Jaeger and E. J. Norcliffe. The Cross-linguistic Study of Sentence Production. *Language and Linguistics Compass*, 3(4):866–887, 2009.

T. F. Jaeger and H. Tily. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335, 2011.

D. Jurafsky. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194, 1996.

D. Jurafsky. Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In *Probabilistic Linguistics*, pages 39–96. MIT Press, 2003.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. 2019. URL `https://web.stanford.edu/~jurafsky/slp3/`.

D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254, 2001.

E. L. Keenan and B. Comrie. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99, 1977.

G. Kempen and K. Harbusch. A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann and C. Habel, editors, *Multidisciplinary Approaches to Language Production*, volume 157, pages 173–182. Mouton de Gruyter, 2004.

P. Kiparsky. The shift to head-initial VP in Germanic. In H. Thráinsson, S. D. Epstein, and S. Peter, editors, *Studies in Comparative Germanic Syntax*, volume 2, pages 140–179. Kluwer Dordrecht, 1996.

S. Kirby. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, 2001.

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

S. Kirby, M. Dowman, and T. L. Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245, 2007.

S. Kirby, T. Griffiths, and K. Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

J. Kizach. Evidence for weight effects in Russian. *Russian linguistics*, 36(3):251–270, 2012.

L. Konieczny. Locality and Parsing Complexity. *Journal of Psycholinguistic Research*, 29 (6):627–645, Nov 2000.

L. Konieczny and P. Döring. Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of the ICCS/ASCS Joint International Conference on Cognitive Science*, pages 13–17, 2003.

H. Kučera and W. N. Francis. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

G. R. Kuperberg and T. F. Jaeger. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59, 2016.

M. Kutas, K. A. DeLong, and N. J. Smith. A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the Brain: Using Our Past to Generate a Future*, 190207, 2011.

W. Labov. *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*, volume 36. John Wiley & Sons, 2011.

R. W. Langacker. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford University Press, 1987.

L. C. Lawyer. The oblique phrase and the order of the relative construction. *STUF-Language Typology and Universals*, 68(4):515–550, 2015.

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

N. Levshina. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help+(to) infinitive in varieties of web-based English. *Glossa: a journal of general linguistics*, 3(1), 2018.

N. Levshina. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572, 2019.

O. Levy and Y. Goldberg. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014a. Association for Computational Linguistics.

O. Levy and Y. Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan, June 2014b. Association for Computational Linguistics.

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

R. Levy. Memory and Surprisal in Human Sentence Comprehension. In R. P. G. van Gompel, editor, *Sentence Processing*, page 78–114. Hove: Psychology Press, 2013.

R. Levy and T. F. Jaeger. Speakers Optimize Information Density Through Syntactic Reduction. In *Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS)*, 2007.

R. Levy, E. Fedorenko, M. Breen, and E. Gibson. The processing of extraposed structures in English. *Cognition*, 122(1):12–36, 2012.

R. Levy, E. Fedorenko, and E. Gibson. The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, 69(4):461–495, 2013.

C. N. Li and S. A. Thompson. Co-verbs in Mandarin Chinese: Verbs or prepositions? *Journal of Chinese Linguistics*, 2(3):257–278, 1974.

T. Linzen. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e98–e108, 2019.

T. Linzen, E. Dupoux, and Y. Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

H. Liu. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191, 2008.

H. Liu. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578, 2010.

H. Liu, Y. Zhao, and W. Li. Chinese Syntactic and Typological Properties Based on Dependency Syntactic Treebanks. *Poznań Studies in Contemporary Linguistics*, 45(4): 509–523, 2009.

S. Liu, H. Mei, A. Williams, and R. Cotterell. On the Idiosyncrasies of the Mandarin Chinese classifier system. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4100–4106, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

A. Lohmann and T. Takada. Order in NP conjuncts in spoken English and Japanese. *Lingua*, 152:48–64, 2014.

B. Lohse, J. A. Hawkins, and T. Wasow. Domain Minimization in English Verb-Particle Constructions. *Language*, 80(2):238–261, 2004.

S. Lubetich and K. Sagae. Data-driven Measurement of Child Language Development with Simple Syntactic Templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics.

M. C. MacDonald. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9(2):157–201, 1994.

M. C. MacDonald. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4:226, 2013.

M. C. MacDonald, N. J. Pearlmutter, and M. S. Seidenberg. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676, 1994.

B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk. Transcription Format and Programs*, volume 1. Psychology Press, 2000.

B. MacWhinney, E. Bates, and R. Kliegl. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23(2): 127–150, 1984.

B. MacWhinney, A. Malchukov, and E. Moravcsik. *Competing Motivations in Grammar and Usage*. Oxford University Press, Oxford, 2014.

K. L. Marblestone. *Semantic and syntactic effects on double prepositional phrase ordering across the lifespan*. PhD thesis, University of Southern California, 2007.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

A. Martin, T. Ratitamkul, K. Abels, D. Adger, and J. Culbertson. Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, 2019.

J. E. Martin. Some competence-process relationships in noun phrases with prenominal and postnominal adjectives. *Journal of Verbal Learning and Verbal Behavior*, 8(4): 471–480, 1969.

R. Marvin and T. Linzen. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

K. Mei. Is modern Chinese really a SOV language? *Cahiers de Linguistique-Asie Orientale*, 7(1):23–45, 1980.

P. Merlo. Predicting word order universals. *Journal of Language Modelling*, 3(2):317–344, 2015.

P. Merlo and E. E. Ferrer. The Notion of Argument in Prepositional Phrase Attachment. *Computational Linguistics*, 32(3):341–378, 2006.

T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

S. Momma, L. R. Slevc, and C. Phillips. The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(5):813, 2016.

E. Morgan and R. Levy. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci)*, page 1649–1654, 2015.

E. Morgan and R. Levy. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:384 – 402, 2016a.

E. Morgan and R. Levy. Frequency-Dependent Regularization in Iterated Learning. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, 2016b.

S. Namboodiripad. A gradient approach to flexible constituent order. 2019. URL `https://doi.org/10.31234/osf.io/rvjn5`.

L. Newbold. Word Order in Hupa. In S. Tuttle and J. Spence, editors, *Working Papers in Athabaskan Languages 2009*, pages 89–106. Alaska Native Language Center Working Papers 8, Fairbanks: Alaska Native Language Center, 2010.

H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, and A. Korhonen. Survey on the Use of Typological Information in Natural Language Processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee.

T. Osborne and K. Gerdes. The status of function words in dependency grammar: A

critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, 4(1), 2019.

R. Östling. Word Order Typology through Multilingual Word Alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China, July 2015. Association for Computational Linguistics.

Y. A. Park and R. Levy. Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 335–343, Boulder, Colorado, June 2009. Association for Computational Linguistics.

J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

M. Peter, F. Chang, J. M. Pine, R. Blything, and C. F. Rowland. When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language*, 81:1–15, 2015.

M. Polinsky. Headedness, again. *Theories of Everything. In Honor of Ed Keenan*, 2012.

C. Pollard and I. A. Sag. *Head-driven Phrase Structure Grammar*. University of Chicago Press, 1994.

E. M. Ponti, H. O'horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601, 2019.

M. Prat-Sala and H. P. Branigan. Discourse Constraints on Syntactic Processing in Language Production: A Cross-Linguistic Study in English and Spanish. *Journal of Memory and Language*, 42(2):168 – 182, 2000.

E. F. Prince. On pragmatic change: The borrowing of discourse functions. *Journal of pragmatics*, 12(5):505–518, 1988.

E. Prud'hommeaux, E. Morley, M. Rouhizadeh, L. Silverman, J. van Santeny, B. Roarkz, R. Sproatz, S. Kauper, and R. DeLaHunta. Computational analysis of trajectories of linguistic development in autism. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–271. IEEE, 2014.

R. Quirk, S. Greenbaum, and G. Leech. *A Comprehensive Grammar of the English Language*. Longman, London, 1985.

R. Rajkumar and M. White. Better surface realization through psycholinguistics. *Language and Linguistics Compass*, 8(10):428–448, 2014.

R. Rajkumar, M. van Schijndel, M. White, and W. Schuler. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, 155:204–232, 2016.

M. Rasekh-Mahand, M. Alizadeh-Sahraie, and R. Izadifar. A corpus-based analysis of relative clause extraposition in Persian. *Ampersand*, 3:21–31, 2016.

F. Reali and M. H. Christiansen. On the necessity of an interdisciplinary approach to language universals. *Language Universals*, pages 266–77, 2009.

F. Reali and T. L. Griffiths. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–328, 2009.

I. Ros. *Minimizing dependencies across languages and speakers. Evidence from Basque, Polish and Spanish and native and non-native bilinguals*. PhD thesis, University of the Basque Country, 2018.

S. Sahakian and B. Snyder. Automatically Learning Measures of Child Language Development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 95–99, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

S. Samonte and G. Scontras. Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. *Proceedings of the Linguistic Society of America*, 4 (1):33–1, 2019.

E. Sapir and V. Golla. Hupa texts, with notes and lexicon. *The Collected Works of Edward Sapir, ed. by Victor Golla & Sean O'Neill*, 14:19–1011, 2001.

M. Schilk and S. Schaub. Noun phrase complexity across varieties of English: Focus on syntactic function and text type. *English World-Wide*, 37(1):58–85, 2016.

S. Schuster and C. D. Manning. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

H. Schütze. Word space. In *Advances in neural information processing systems*, pages 895–902, 1993.

G. Scontras, J. Degen, and N. D. Goodman. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66, 2017.

G. Scontras, J. Degen, and N. D. Goodman. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*, 12, 2019.

A. Siewierska, editor. *Constituent Order in the Languages of Europe*. Walter de Gruyter Co., 1998.

K. Smith, A. Perfors, O. Fehér, A. Samara, K. Swoboda, and E. Wonnacott. Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160051, 2017.

E. Spair. Language: An introduction to the study of speech. *NY: Harcourt, Brace & Co.*, 1921.

J. Spence. Some Discourse Effects on Split Nominals in Hupa, 2008.

J. Spence. *Language Change, Contact, and Koineization in Pacific Coast Athabaskan*. Doctoral dissertation, University of California, Berkeley, 2013.

R. Sproat and C. Shih. The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos and R. Ishihara, editors, *Interdisciplinary Approaches to Language*, pages 565–593. Springer, 1991.

L. M. Stallings, M. C. MacDonald, and P. G. O'Seaghdha. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417, 1998.

A. Staub and C. J. Clifton. Syntactic prediction in language comprehension: Evidence from either... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2):425, 2006.

M. Straka and J. Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.

C.-F. Sun and T. Givón. On the so-called SOV word order in Mandarin Chinese: A quantified text study and its implications. *Language*, pages 329–351, 1985.

D. Temperley. Minimization of dependency length in written English. *Cognition*, 105(2): 300–333, 2007.

D. Temperley and D. Gildea. Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal? *Annual Review of Linguistics*, 4(1):67–80, 2018.

L. Tesnière. *Éléments de syntaxe structurale, Klincksieck.* Pans, 1959.

R. Thornton, M. Gil, and M. C. MacDonald. Accounting for crosslinguistic variation: A constraint-based perspective. *Syntax and Semantics*, pages 211–227, 1998.

H. Tily. *The Role of Processing Complexity in Word Order Variation and Change.* PhD thesis, Stanford University, 2010.

R. S. Tomlin. *Basic Word Order: Functional Principles.* London: Croom Helm, 1986.

A. Trotzke and E. Wittenberg. Long-standing issues in adjective order and corpus evidence for a multifactorial approach. *Linguistics*, 57(2):273–282, 2019.

J. C. Trueswell. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35(4):566–585, 1996.

J. C. Trueswell, M. K. Tanenhaus, and C. Kello. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3):528, 1993.

T. Tsunoda, S. Ueda, and Y. Itoh. Adpositions in word-order typology. *Linguistics*, 33 (4):741–762, 1995.

M. Ueno and M. Polinsky. Does headedness affect processing? A new look at the VO–OV contrast. *Journal of Linguistics*, 45(3):675–710, 2009.

M. van Schijndel, A. Mueller, and T. Linzen. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

S. Vasishth and R. L. Lewis. Argument-Head Distance and Processing Complexity: Explaining Both Locality and Antilocality Effects. *Language*, 82(4):767–794, 2006.

D. Wang and J. Eisner. Fine-Grained Prediction of Syntactic Typology: Discovering Latent Structure with Supervised Learning. *Transactions of the Association for Computational Linguistics (TACL)*, 5:147–161, June 2017.

D. Wang and J. Eisner. Synthetic data made to order: The case of parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1325–1337, Brussels, Nov. 2018a.

D. Wang and J. Eisner. Surface statistics of an unknown language indicate how to parse it. *Transactions of the Association for Computational Linguistics (TACL)*, 6:667–685, Dec. 2018b.

W. S.-Y. Wang and C. Sun. *The Oxford Handbook of Chinese Linguistics*. Oxford Handbooks, 2015.

T. Wasow. End-Weight from the Speaker's Perspective. *Journal of Psycholinguistic Research*, (3):347–361, 1997a.

T. Wasow. Remarks on grammatical weight. *Language Variation and Change*, (1):81–105, 1997b.

T. Wasow and J. Arnold. Post-verbal constituent ordering in English. *Topics in English Linguistics*, 43:119–154, 2003.

T. Wasow, T. F. Jaeger, and D. Orr. Lexical variation in relativizer frequency. *Expecting the Unexpected: Exceptions in Grammar*, pages 175–96, 2011.

M. White and R. Rajkumar. Minimal Dependency Length in Realization Ranking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 244–255, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

D. Wiechmann and A. Lohmann. Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, 25(1):65–88, 2013.

E. Wilcox, P. Qian, R. Futrell, M. Ballesteros, and R. Levy. Structural Supervision Improves Learning of Non-local Grammatical Dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3302–3312, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

S. Wulff. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8(2):245–282, 2003.

S. Wulff and S. T. Gries. Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, 5(1): 122–150, 2015.

N. Xue, F. Xia, F.-D. Chiou, and M. Palmer. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238, 2005.

H. Yamashita. Scrambled sentences in Japanese: Linguistic properties and motivations for production. *Text – Interdisciplinary Journal for the Study of Discourse*, 22(4):597–634, 2002.

H. Yamashita and F. Chang. "Long before short" preference in the production of a head-final language. *Cognition*, 81(2):B45–B55, 2001.

E. Yi, J.-P. Koenig, and D. Roland. Semantic similarity to high-frequency verbs affects syntactic frame selection. *Cognitive Linguistics*, 30(3):601–628, 2019.

D. Zeman, J. Nivre, M. Abrams, N. Aepli, Ž. Agić, L. Ahrenberg, G. Aleksandravičiūtė, L. Antonsen, K. Aplonova, M. J. Aranzabe, G. Arutie, M. Asahara, L. Ateyah, M. Attia, A. Atutxa, L. Augustinus, E. Badmaeva, M. Ballesteros, E. Banerjee, S. Bank,

V. Barbu Mititelu, V. Basmov, C. Batchelor, J. Bauer, S. Bellato, K. Bengoetxea, Y. Berzak, I. A. Bhat, R. A. Bhat, E. Biagetti, E. Bick, A. Bielinskienė, R. Blokland, V. Bobicev, L. Boizou, E. Borges Völker, C. Börstell, C. Bosco, G. Bouma, S. Bowman, A. Boyd, K. Brokaitė, A. Burchardt, M. Candito, B. Caron, G. Caron, T. Cavalcanti, G. Cebiroğlu Eryiğit, F. M. Cecchini, G. G. A. Celano, S. Čéplö, S. Cetin, F. Chalub, J. Choi, Y. Cho, J. Chun, A. T. Cignarella, S. Cinková, A. Collomb, Ç. Çöltekin, M. Connor, M. Courtin, E. Davidson, M.-C. de Marneffe, V. de Paiva, E. de Souza, A. Diaz de Ilarraza, C. Dickerson, B. Dione, P. Dirix, K. Dobrovoljc, T. Dozat, K. Droganova, P. Dwivedi, H. Eckhoff, M. Eli, A. Elkahky, B. Ephrem, O. Erina, T. Erjavec, A. Etienne, W. Evelyn, R. Farkas, H. Fernandez Alcalde, J. Foster, C. Freitas, K. Fujita, K. Gajdošová, D. Galbraith, M. Garcia, M. Gärdenfors, S. Garza, K. Gerdes, F. Ginter, I. Goenaga, K. Gojenola, M. Gökırmak, Y. Goldberg, X. Gómez Guinovart, B. González Saavedra, B. Griciūtė, M. Grioni, N. Grūzītis, B. Guillaume, C. Guillot-Barbance, N. Habash, J. Hajič, J. Hajič jr., M. Hämäläinen, L. Hà Mỹ, N.-R. Han, K. Harris, D. Haug, J. Heinecke, F. Hennig, B. Hladká, J. Hlaváčová, F. Hociung, P. Hohle, J. Hwang, T. Ikeda, R. Ion, E. Irimia, Ọ. Ishola, T. Jelínek, A. Johannsen, F. Jørgensen, M. Juutinen, H. Kaşıkara, A. Kaasen, N. Kabaeva, S. Kahane, H. Kanayama, J. Kanerva, B. Katz, T. Kayadelen, J. Kenney, V. Kettnerová, J. Kirchner, E. Klementieva, A. Köhn, K. Kopacewicz, N. Kotsyba, J. Kovalevskaitė, S. Krek, S. Kwak, V. Laippala, L. Lambertino, L. Lam, T. Lando, S. D. Larasati, A. Lavrentiev, J. Lee, P. Lê Hồng, A. Lenci, S. Lertpradit, H. Leung, C. Y. Li, J. Li, K. Li, K. Lim, M. Liovina, Y. Li, N. Ljubešić, O. Loginova, O. Lyashevskaya, T. Lynn, V. Macketanz, A. Makazhanov, M. Mandl, C. Manning, R. Manurung, C. Mărănduc, D. Mareček, K. Marheinecke, H. Martínez Alonso, A. Martins, J. Mašek, Y. Matsumoto, R. McDonald, S. McGuinness, G. Mendonça, N. Miekka, M. Misirpashayeva, A. Missilä, C. Mititelu, M. Mitrofan, Y. Miyao, S. Montemagni, A. More, L. Moreno Romero, K. S. Mori, T. Morioka, S. Mori, S. Moro, B. Mortensen, B. Moskalevskyi, K. Muischnek, R. Munro, Y. Murawaki, K. Müürisep, P. Nainwani, J. I. Navarro Horñiacek, A. Nedoluzhko, G. Nešpore-Bērzkalne, L. Nguyễn Thị, H. Nguyễn Thị Minh, Y. Nikaido, V. Nikolaev,

R. Nitisaroj, H. Nurmi, S. Ojala, A. K. Ojha, A. Olúòkun, M. Omura, P. Osenova, R. Östling, L. Øvrelid, N. Partanen, E. Pascual, M. Passarotti, A. Patejuk, G. Paulino-Passos, A. Peljak-Łapińska, S. Peng, C.-A. Perez, G. Perrier, D. Petrova, S. Petrov, J. Phelan, J. Piitulainen, T. A. Pirinen, E. Pitler, B. Plank, T. Poibeau, L. Ponomareva, M. Popel, L. Pretkalniņa, S. Prévost, P. Prokopidis, A. Przepiórkowski, T. Puolakainen, S. Pyysalo, P. Qi, A. Rääbis, A. Rademaker, L. Ramasamy, T. Rama, C. Ramisch, V. Ravishankar, L. Real, S. Reddy, G. Rehm, I. Riabov, M. Rießler, E. Rimkutė, L. Rinaldi, L. Rituma, L. Rocha, M. Romanenko, R. Rosa, D. Rovati, V. Roșca, O. Rudina, J. Rueter, S. Sadde, B. Sagot, S. Saleh, A. Salomoni, T. Samardžić, S. Samson, M. Sanguinetti, D. Särg, B. Saulīte, Y. Sawanakunanon, N. Schneider, S. Schuster, D. Seddah, W. Seeker, M. Seraji, M. Shen, A. Shimada, H. Shirasu, M. Shohibussirri, D. Sichinava, A. Silveira, N. Silveira, M. Simi, R. Simionescu, K. Simkó, M. Šimková, K. Simov, A. Smith, I. Soares-Bastos, C. Spadine, A. Stella, M. Straka, J. Strnadová, A. Suhr, U. Sulubacak, S. Suzuki, Z. Szántó, D. Taji, Y. Takahashi, F. Tamburini, T. Tanaka, I. Tellier, G. Thomas, L. Torga, T. Trosterud, A. Trukhina, R. Tsarfaty, F. Tyers, S. Uematsu, Z. Urešová, L. Uria, H. Uszkoreit, A. Utka, S. Vajjala, D. van Niekerk, G. van Noord, V. Varga, E. Villemonte de la Clergerie, V. Vincze, L. Wallin, A. Walsh, J. X. Wang, J. N. Washington, M. Wendt, S. Williams, M. Wirén, C. Wittern, T. Woldemariam, T.-s. Wong, A. Wróblewska, M. Yako, N. Yamazaki, C. Yan, K. Yasuoka, M. M. Yavrumyan, Z. Yu, Z. Žabokrtský, A. Zeldes, M. Zhang, and H. Zhu. Universal Dependencies 2.5, 2019. URL `http://hdl.handle.net/11234/1-3105`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

N. N. Zhang. Adpositions. In R. Sybesma, W. B. Y. Gu, Z. Handel, and C.-T. J. H. J. Myers, editors, *Encyclopedia of Chinese Language and Linguistics*, page 116–122. Brill, 2013.