Assessing Corpus Evidence for Formal and Psycholinguistic Constraints on Nonprojectivity

Himanshu Yadav
Department of Linguistics
University of Potsdam
hyadav@uni-potsdam.de

Samar Husain
Department of Humanities and
Social Sciences
Indian Institute of Technology, Delhi
samar@hss.iitd.ac.in

Richard Futrell
Department of Language Science
University of California, Irvine
rfutrell@uci.edu

Formal constraints on crossing dependencies have played a large role in research on the formal complexity of natural language grammars and parsing. Here we ask whether the apparent evidence for constraints on crossing dependencies in treebanks might arise because of independent constraints on trees, such as low arity and dependency length minimization. We address this question using two sets of experiments. In Experiment 1, we compare the distribution of formal properties of crossing dependencies, such as gap degree, between real trees and baseline trees matched for rate of crossing dependencies and various other properties. In Experiment 2, we model whether two dependencies cross, given certain psycholinguistic properties of the dependencies. We find surprisingly weak evidence for constraints originating from the mild context-sensitivity literature (gap degree and well-nestedness) beyond what can be explained by constraints on rate of crossing dependencies, topological properties of the trees, and dependency length. However, measures that have emerged from the parsing literature (e.g., edge degree, endpoint crossings, and heads' depth difference) differ strongly between real and random trees. Modeling results show that cognitive metrics relating to information locality and working-memory limitations affect whether two dependencies cross or not, but they do not fully explain the distribution of crossing dependencies in natural languages. Together these results suggest

Submission received: 2 May 2021; revised version received: 13 December 2021; accepted for publication: 29 December 2021.

https://doi.org/10.1162/coli_a_00437

that crossing constraints are better characterized by processing pressures than by mildly context-sensitive constraints.

1. Introduction

The syntactic structure of natural language sentences can be captured to a large extent using **dependency trees**: directed trees drawn over words indicating what words are dependent on what other words (Tesnière 1959; Hays 1964; Mel'čuk 1988; Hudson 1990; Nivre 2015). An example is shown in Figure 1. A number of key formal questions in linguistics boil down to questions about the structure of these dependency trees. In particular, recent work has concluded that the characterization of natural language in formal language theory depends on the constraints that can be placed on crossing dependencies in dependency trees (Kuhlmann 2007).

Here we use recently available massively crosslinguistic dependency treebanks (Nivre et al. 2015; Gerdes et al. 2018, 2019) to take up the question, what distinguishes natural language dependency trees within the space of all possible tree structures, in terms of crossing dependencies? We investigate two kinds of proposed constraints on dependency trees. First, we investigate a set of graph-theoretic constraints originating in the formal language theory and dependency parsing literatures, which have bearing on the location of natural language within the Chomsky hierarchy. Comparing natural language dependency trees to random trees of various kinds, we find little evidence that linguistic trees are constrained by the kinds of formal properties studied in formal language theory literature, but we do find strong evidence for under-studied constraints originating from the parsing literature. Second, we explore a set of performance-based and psycholinguistically motivated soft constraints, motivated in terms of empirically measured human online processing difficulty, finding evidence that these constrain crossing dependencies.

1.1 Background

The attempt to characterize the complexity of natural language in terms of formal language theory has been an extraordinarily productive enterprise joining linguistics, computer science, and mathematics (Chomsky 1956; Chomsky and Schützenberger 1963; Hopcroft and Ullman 1979). In recent decades, a consensus has emerged that the syntactic structure of natural languages is well characterized in terms of the **mildly context-sensitive hierarchy** of languages (Weir 1988; Joshi, Shanker, and Weir 1991; Michaelis 1998; Kuhlmann 2013), a complexity class lying between context-free and context-sensitive and characterized by formal restrictions on various kinds of discontinuity in constituents. In dependency frameworks, these discontinuous constituents correspond to crossing dependencies (see Figure 1 for a simple example). Therefore, formal restrictions on discontinuous constituents correspond to formal restrictions on crossing dependencies in the ordered dependency tree (Kuhlmann 2013).

A number of formal restrictions on crossing dependencies have been proposed in the last 20 years, going beyond the simple observation that crossing dependencies are rare (Havelka 2007; Ferrer-i-Cancho, Gómez-Rodríguez, and Esteban 2018). We call these formal constraints on crossing dependencies **crossing constraints**. For example, Kuhlmann (2013) has proposed that dependency trees have limited gap degree and are usually well-nested (see Figure 2b). Pitler, Kannan, and Marcus (2013) propose that crossing dependency configurations have a property called 1-end-point-crossing. Other

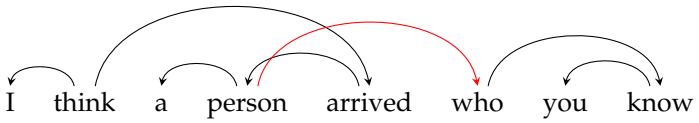


Figure 1
An example dependency tree. Arrows point from heads to dependents. This tree has one crossing dependency, marked in red.

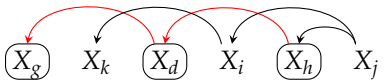


Figure 2
The gap degree of the tree above is 2.

formal restrictions such as edge degree, multiplanarity, and heads' depth difference have also been proposed (Yli-Jyrä 2003; Kuhlmann and Nivre 2006; Nivre 2007; Yadav, Vaidya, and Husain 2017). Among these crossing constraints, the constraint on gap degree is especially interesting, because gap degree defines the position of a formal language in the mildly context-sensitive hierarchy: A mildly context-sensitive language is defined by a finite upper bound on gap degree, with gap degree zero corresponding to a context-free grammar (Kuhlmann 2013; Marcus 1965).

The empirical arguments that crossing dependencies are constrained by factors such as gap-degree, and so forth, have typically come from demonstrations that crossing dependencies in a large number of observed trees in dependency corpora violate the constraints never or rarely (e.g., Kuhlmann and Nivre 2006; Havelka 2007). For example, Kuhlmann and Nivre (2006) show that only 0.17% of analyses in a Danish Dependency Treebank have gap-degree 2 and 99% of all non-projective structures are well-nested. These findings do not rule out the possibility that crossing constraints might manifest as epiphenomena of other, more general properties of dependency trees.

To appreciate this possibility, consider gap degree as an example. Gap degree is defined as the number of discontinuities in the projection of a node (see Figure 2 for an example), so it is upper-bounded by the number of discontinuities in the sentence. Given that crossing dependencies / discontinuous constituents are rare (Ferrer-i-Cancho, Gómez-Rodríguez, and Esteban 2018), we might expect to measure a low gap degree even if there is no true causally active constraint against gap degree.

This observation opens the possibility that crossing constraints such as gap degree, edge degree, and so on, could be epiphenomena of generic graph-theoretic properties of dependency trees, such as their height, arity, and so forth, and that together these generic factors drive the observable formal properties of crossing dependencies in natural language. If true, this would mean that apparent bounds on gap degree are accidental, and that formalisms such as mildly context-sensitive grammars fit linguistic data only because the structures that would violate them are rare by chance. It is this possibility that we explore in depth in Section 2.

Formal constraints such as mild context-sensitivity are usually associated with a competence-based approach to defining the generative capacity of language, where it

is posited that there are (possibly innate) formal constraints on possible mental grammars. Such an approach can be evaluated against a performance-based approach. For example, a well-known cross-linguistic phenomenon termed **syntactic islands** is typically explained via constraints on structural configurations (Chomsky 1981). Broadly construed, "islands" prohibit formation of certain crossing dependencies and hence are important in circumscribing the formal domain of natural language. Notwithstanding the competence-based explanation, it has been argued that island constraints arise not due to structural ill-formedness, but rather due to constraints on human online language processing (e.g., Hofmeister and Sag 2010).

In a similar vein, the distribution of crossing dependencies in natural languages could also be understood in terms of processing-related constraints. It is a well known fact that, cross-linguistically, simple linguistic codes are more frequent, while complex linguistic codes are rare (e.g., Zipf 1949; Mahowald et al. 2013; Piantadosi 2014; Ferrer-i-Cancho et al. 2013; Ferrer-i-Cancho, Bentz, and Seguin 2020). Such a pattern has been argued to highlight the communicative efficiency of natural language (Jaeger and Tily 2011; Gibson et al. 2019). On this account, the restrictions on crossing dependencies in natural language could arise because such syntactic configurations are difficult (but not impossible) to produce and comprehend (Bresnan et al. 1982; Ades and Steedman 1982; Bach, Brown, and Marslen-wilson 1986; Joshi 1990; Ferrer-i-Cancho 2006, 2014; Gómez-Rodríguez and Ferrer-i-Cancho 2017; Gómez-Rodríguez, Christiansen, and Ferrer-i-Cancho 2019). For example, given the incremental nature of language production (e.g., Ferreira and Henderson 1998), it could be assumed that production of crossing dependencies, which necessarily involves a discontinuity in a phrasal boundary, would incur an increased processing cost compared with non-crossing dependencies. Although there has been some experimental work on investigating the processing cost incurred during comprehension of sentences with crossing dependencies (e.g., Levy et al. 2012; Staub et al. 2018), a corpus-based empirical study investigating the influence of psycholinguistic factors beyond dependency distance on crossing dependencies is lacking. If certain psycholinguistic factors (e.g., working-memory constraints) can explain the occurrence of crossing dependencies, that could suggest a functional motivation of why such dependencies are rare in natural language.

In this article, we conduct a cross-linguistic corpus investigation into crossing dependencies both from the perspective of formal crossing constraints as well as from the perspective of processing constraints. In the first set of experiments we investigate if well-known crossing constraints (e.g., gap-degree, edge-degree) can account for crossing dependencies attested in various dependency treebanks. In particular, Section 2 compares the formal properties of crossing dependencies in real dependency trees with various random baselines matched in the number of crossing dependencies and other global graph-theoretic factors. In the second set of experiments, we conduct preliminary analyses to investigate whether certain psycholinguistic factors can account for crossing dependencies attested in various dependency treebanks. In particular, Section 3 models the tendency of two dependency arcs to cross given certain psycholinguistic metrics that are computed using the local configuration of the two arcs.

2. Global Graph-Theoretic Factors

In this section we investigate evidence for global graph-theoretic constraints on crossing dependencies originating from the literatures on formal language theory and dependency parsing. Our goal is to determine which, if any, of these constraints really distinguish natural language dependency trees from the space of all possible directed

trees. To do so, we study how often these contraints are violated in (1) real dependency trees and (2) a collection of random baseline trees controlling various properties of trees.

2.1 Background

Chomsky (1956, 1957) first posed the question of how to characterize grammars of natural languages as computational objects, launching a research program that has attempted to describe sets of grammatical sentences using tools from mathematical logic, graph theory, and automaton theory. Since the late 1980s, a consensus has emerged that natural language is well characterized as falling within the **mildly context-sensitive** class of languages, a formal language class that is larger than the context-free languages, yet without taking advantage of the full expressive power of context-sensitive languages (Weir 1988; Joshi, Shanker, and Weir 1991).

These mildly context-sensitive languages are defined by constraints which turn out to be equivalent to constraints on crossing dependencies. Therefore, by answering the question of what constrains crossing dependencies, we can make progress toward understanding human languages at a computational level.

Crossing dependencies are related to formal language theory because they correspond to **displacement** phenomena in languages—structures that cannot be captured by a context-free grammar. Across grammatical formalisms, displacement phenomena and are modeled using a distinct kind of structure from non-crossing dependencies. Displacement phenomena (encompassing both extraposition and *wh*-dependencies) have been modeled in various ways in different syntactic frameworks:

- In the Minimalist tradition, non-crossing dependencies correspond to structures that can be built by the computational operation MERGE (or "external merge"), while crossing dependencies arise from the action of a distinct structure-building operation MOVE (or "internal merge") (Chomsky 1995; Stabler 1997; Michaelis 1998). A grammar with only MERGE would generate context-free languages and projective dependency trees.
- In phrase structure-based frameworks such as Lexical Functional Grammar (Bresnan 1982), Head-driven Phrase Structure Grammar (Pollard and Sag 1994), and Combinatory Categorial Grammar (Steedman and Baldridge 2011), displacement phenomena are handled using phrase structure rules defined in a way that allows information to percolate through a tree in a non-local manner, a mechanism called "slash-passing."
- Some theories of dependency grammar invoke the idea that each word in a sentence has both a syntactic "head" and a syntactic "governor," which may coincide. Arcs drawn from governors to dependents may cross, but arcs drawn from heads to dependents never cross (Groß and Osborne 2009). Crossing dependencies correspond to cases where the syntactic governor is distinct from the syntactic head. What we are calling "heads" in this article would be "governors" in such theories.

These various formalisms allow for crossing dependencies by different mechanisms, yet with constraints which turn out to be similar or equivalent across formalisms. In particular, most mildly context-sensitive formalisms end up instantiating bounds on a quantity called **gap degree**. This quantity goes by different names depending on

the grammar formalism. It is equivalent (up to additive constants) to *block degree* or *fan-out* in linear context-free rewriting systems (Kuhlmann 2007, 2013), the number of *components* in multiple context-free grammars (Seki et al. 1991), the maximal *rank* of a coupled context-free grammar (Hotz and Pitsch 1996), the number of *licensee features* in Minimalist Grammars (Michaelis 1998; Boston, Hale, and Kuhlmann 2010), and others. It was first introduced in a dependency framework by Holan et al. (1998), and shown to relate to mild context-sensitivity by Kuhlmann (2007). Some mildly context-sensitive formalisms also induce a constraint called **well-nestedness**, which can also be reduced to constraints on crossing dependencies (Bodirsky, Kuhlmann, and Möhl 2005).

Crossing constraints have also been of interest for those studying the development of efficient dependency parsing algorithms. Such algorithms are generally only available for trees with constrained crossings. For example, if we assume that all trees are projective, then we can perform exact parsing in time cubic in the sentence length ($O(n^3)$) by reducing the dependency grammar to a lexicalized context-free grammar (Eisner and Satta 1999). If we assume all trees are well-nested and gap degree is bounded, then we can generally parse in polynomial time. Without the constraint of well-nestedness, parsing becomes NP-hard (Satta 1992; Gómez-Rodríguez, Carroll, and Weir 2011).

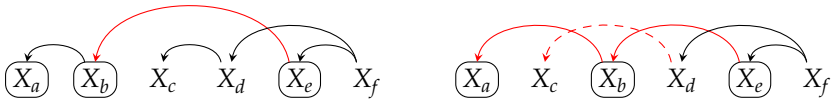
The parsing literature has also been the source of a number of new formal constraints on crossing dependencies, beyond those introduced in the formal syntax literature. For example, Pitler, Kannan, and Marcus (2013) propose a constraint called 1-end-point-crossing. If we assume that all dependency trees are 1-end-point-crossing, then we can parse in quartic time ($O(n^4)$) (see also Gómez-Rodríguez, Shi, and Lee 2018).

2.2 Constraints Considered

Our goal is to determine if there is really evidence for formal graph-theoretic crossing constraints on crossing dependencies in depedency treebanks beyond what can be explained in terms of more generic properties of dependency trees. Below, we list and define the formal crossing constraints that we test. In our terminology we strive to follow Kuhlmann and Nivre (2006).

Gap degree. The **projection** of a node X is the ordered list of all the nodes transitively dominated by X plus X itself. For example, in the dependency tree in Figure 2, $\left[X_g, X_d, X_h\right]$ is the projection of the node X_h. A projection is discontinuous if it forms a discontinuous substring of the sentence. For example, the projection of X_h has two discontinuities, one between X_d and X_h, and another between X_g and X_d. The **gap degree** of a tree is the largest number of discontinuities in the projection of any node in the tree.

Well-nestedness. The **subtree** rooted at a node X is the set of all the transitive nodes dominated by X plus X itself. For example, in the dependency tree (a) and (b) in Figure 3, $\{X_a, X_b, X_e\}$ is the subtree rooted at node X_e, and $\{X_c, X_d\}$ is the subtree rooted at node X_d. Two subtrees with nodes $\{P, Q\}$ and $\{R, S\}$ **interleave** if the nodes are in linear order such that $P < R < Q < S$. A dependency tree is **ill-nested** if and only if two of its disjoint subtrees interleave. For example, in (3a), $\{X_a, X_b, X_e\}$ and $\{X_c, X_d\}$ are two disjoint subtrees but they do not interleave as the nodes are in the order $X_a < X_b < X_c < X_d < X_e$. Therefore, tree (3a) is well-nested. In (3b), the disjoint subtrees $\{X_a, X_b, X_e\}$ and $\{X_c, X_d\}$ interleave as the order of the nodes is $X_a < X_c < X_b < X_d < X_e$. The dashed red arc creates the ill-nestedness. Ill-nestedness implies gap degree > 1.



(a) Well-nested, Gap degree 1 (b) Ill-nested, Gap degree 2

Figure 3
Schematic of well-nested and ill-nested trees.

(a) Edge degree = 2, End-point crossing = 1 **(b)** Edge degree = 2, End-point crossing = 2

Figure 4
Dependency schemas showing edge degree and end-point crossing. In both the dependency trees, $X_h \rightarrow X_d$ is a crossing dependency. The span of crossing dependency e consists of X_i, X_a, and X_b. Nodes X_i and X_b are dominated neither by head X_h nor by any node in span e. In **(a)** and **(b)**, different sets of nodes are modified by X_i and X_b.

Edge degree. Let e be the span of dependency arc $X_h \rightarrow X_d$. The span e consists of nodes between a head X_h and its dependent X_d, which are X_i, X_a, and X_b in Figure 4. The **edge degree** of a dependency arc $X_h \rightarrow X_d$ is the number of nodes in the span e that are neither transitively dominated by some node in the span e nor transitively dominated by the head X_h (Kuhlmann and Nivre 2006). For example, the arc $X_h \rightarrow X_d$ in Figure 4a and 4b has an edge degree of 2 because nodes X_i and X_b are not dominated by any node in the span e. In addition, they are also not dominated by the head X_h. The edge degree of a dependency tree is the highest edge degree among the arcs of the tree.

There are cognitive reasons to suspect that edge degree might be limited in natural language. From an online processing perspective, higher edge degree in a subtree results in a need to maintain an unresolved crossing dependency across a longer span of words, which may result in online processing difficulty due to higher working memory load (Gibson 1998).

End-point crossings. The **end-point crossings** of a dependency arc is the number of distinct heads of all edges that cross the arc. More formally, given an arc $X_h \rightarrow X_d$ with a span e, the end-point crossings of arc $X_h \rightarrow X_d$ is defined as the number of distinct heads of nodes in e that are not part of the projection of X_h nor of any element of e. The end-point crossings of a tree is the maximum end-point crossings of any arc in the tree. For example, in Figure 4a, the number of heads modified by X_i and X_b is 1 (corresponding to X_j), therefore, the end-point crossing is 1. In Figure 4b, the number of heads modified by X_i and X_b are 2 (corresponding to X_j and X_r, respectively); therefore, the end-point crossing is 2.

It has been argued that natural language dependency trees tend to have not more than one end-point crossing, which is called the 1-end-point-crossing constraint. Pitler, Kannan, and Marcus (2013) argue that this constraint is related to the Phase Impenetrability Condition from Minimalist syntax (Chomsky 2007). From a processing based perspective, higher end-point crossings in a subtree should lead to multiple

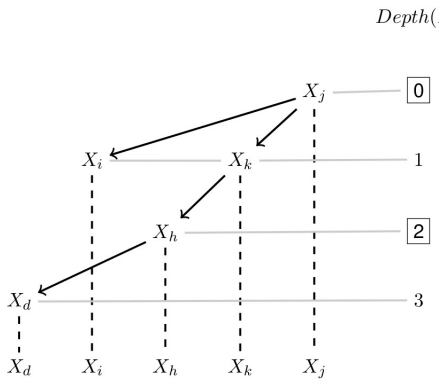


Figure 5
A schematic diagram for heads' depth difference (HDD).

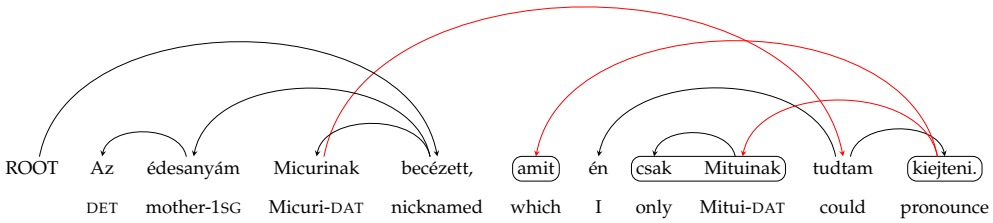
heads/dependents being maintained/stored at the same time in the parse stack. This should lead to increased storage cost (Gibson 1998). In addition, a longer span of the crossing dependency could lead to similarity-based interference (Lewis and Vasishth 2005) at the head.

Heads' depth difference. For a crossing dependency $X_h \rightarrow X_d$, suppose X_i is the node that creates a discontinuity, that is, X_i is not directly or indirectly dominated by X_h (see Figure 5). For this configuration, we call X_i the **intervener**, X_j the head of the intervener, and X_h the head of the crossing dependency. The **heads' depth difference** (HDD) of an arc is defined as the difference between the depth of the head of the crossing dependency X_h and the depth of the head of the intervener X_j. This is schematically shown in Figure 5. Depth of a node is computed as the hierarchical position of that node in a projection chain. The depth of X_h is 2 while the depth of X_j is 0, making the HDD for this configuration equal to 2. Thus, HDD for a crossing dependency $X_h \rightarrow X_d$ is:

$$\text{HDD}(X_h, X_d) = \text{depth}(X_h) - \text{depth}(X_j) \tag{1}$$

where depth(X_h) is the hierarchical position of the head of the non-projective dependency (X_h) and depth(X_j) is the hierarchical position of the head of the intervening element (X_i). The HDD of a dependency tree is the maximum HDD among the HDDs of the arcs in the tree.

In terms of formal syntax, HDD can correspond to the hierarchical depth between a filler and a gap in a long distance dependency (e.g., *wh* movement). Based on the theoretical syntax literature, HDD should be unbounded, at least for leftward *wh*-dependencies (Sag, Wasow, and Bender 1999). However, increasing HDD seems to correlate with increased online processing difficulty for humans (Phillips, Kazanina, and Abada 2005). More generally, HDD has been proposed (see Yadav, Vaidya, and Husain 2017) to formalize the experimental findings that increased embedding depth leads to processing difficulty (e.g., Yngve 1960; Gibson and Thomas 1999). Therefore, it is possible that HDD is restricted in dependency trees due to cognitive constraints.



"My mother nicknamed me Micuri, which I could only pronounce as Mitui."

Figure 6
Dependency tree of a Hungarian sentence from the Szeged Dependency Treebank in UD 2.8 (punctuation removed from the dependency structure). The SUD parse is the same. The projection of the final verb *kiejteni* is shown in boxes.

2.2.1 Example Tree. In order to demonstrate our crossing constraints, Figure 6 shows an example of a complex dependency tree in Hungarian, drawn from the Szeged Dependency Treebank. The tree has 3 crossing dependencies, with gap degree 2: The projection from *kiejteni* has two discontinuities. The tree has edge degree 2, end-point crossings 2, and HDD 2. It has sentence length 10, arity[1] 2, and tree height[2] 6.

2.3 Methods

To test if crossing constraints (e.g., gap degree, edge degree) can account for crossing dependencies attested in natural language treebanks, we compare the distribution of crossing constraints in natural languages with the random baselines matched in number of crossing dependencies and other tree properties. For example, to test if gap degree is a constraint—over and above the constraint on number of crossings and dependency lengths—we compare the distribution of gap degree in natural languages with the random baseline matched in number of crossings and dependency lengths. We describe the random baselines and statistical method used for comparison next (cf. Yadav, Husain, and Futrell 2019).

2.3.1 Random Baselines. We use four random baselines to assess whether crossing constraints occur independent of rate of crossing dependencies and other tree properties like tree height or dependency lengths in natural languages. They include *random trees*, *random linear arrangements*, dependency length (DL)-*controlled random trees*, and *DL-controlled RLAs*. Each baseline controls a particular set of tree properties as shown in Table 1.

In order to generate a **random tree** corresponding to a real language tree of sentence length n, we first sample from a uniform distribution over tree structures with n nodes using Prüfer codes (Prüfer 1918), and then use rejection sampling to obtain a tree that matches the real tree in the number of crossing dependencies. The resulting distribution is uniform over all tree structures with the specified length and number of crossing edges. A DL-controlled random tree must match a real tree both in terms of its number of crossings and in terms of the distribution of dependency lengths within the tree (for

1 We define arity of a tree as maximum of out-degree of its nodes.
2 Tree height is maximum distance from the root to a leaf node of the tree.

Table 1
Properties of random baselines: Each baseline matches in certain properties with the natural language trees. The tree properties controlled in baseline are indicated by ✓.

Baseline	Controlled tree properties		
	Number of crossings	Tree topology	Dependency length
Random trees	✓		
RLAs	✓	✓	
DL-controlled random trees	✓		✓
DL-controlled RLAs	✓	✓	✓

details, see Yadav, Husain, and Futrell 2021). This procedure samples from the uniform distribution on trees with specified length, number of crossings, and distribution of dependency lengths.

Random linear arrangements (RLAs) are generated by permuting the linear order of nodes in a real languages tree. A reordered tree that matches in number of crossings with a real tree is accepted as a valid sample for RLAs. To generate **DL-controlled RLAs**, we sample from RLAs that match in the distribution of dependency lengths with the real trees.

These baseline trees are all generated by a rejection sampling procedure that rejects the vast majority of samples. As such, it is only possible to generate sentences with length up to 11 using this method given currently available computing resources.

All baselines studied in this article control for the number of crossings per sentences. For baselines that control for dependency lengths but not the number of crossings, see Yadav, Husain, and Futrell (2021).

2.3.2 Data. For natural languages data, we use treebanks from Surface-syntactic Universal Dependencies (SUD) v.2.4 (Gerdes et al. 2018, 2019). We test on treebanks from 56 languages, excluding treebanks with less than 500 sentences and ancient languages.

The data used for the analysis contained 9 head-final and 47 head-initial languages for the analysis, as determined by Yadav et al. (2020), which focuses on verb–object relations. The head-final languages were: Afrikaans, Dutch, German, Hindi, Japanese, Korean, Persian, Tamil, and Urdu. The head-initial languages were: Amharic, Arabic, Bulgarian, Bambara, Catalan, Czech, Danish, Greek, English, Spanish, Estonian, Basque, Finnish, Faroese, French, Irish, Galician, Hebrew, Croatian, Upper Sorbian, Hungarian, Armenian, Indonesian, Italian, Kazakh, Northern Kurdish, Lithuanian, Latvian, Maltese, Erzya, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Northern Sami, Serbian, Swedish, Thai, Turkish, Uyghur, Ukrainian, Vietnamese, Wolof, and Chinese.

The SUD treebanks have been converted from the Universal Dependencies (UD) treebanks (Nivre 2015) to reflect syntactic dependencies rather than the more semantic dependencies favored by UD. We also did an additional analysis on the corresponding UD treebanks. We found that the rate of violating crossing constraints is lower in UD trees but the overall pattern of results is the same as for SUD treebanks with two exceptions: (1) the constraint on end-point crossings receives weaker evidence, and (2) the constraint on well-nestedness receives stronger evidence. Regarding point (1), this is likely due to UD's flat structure, which means that many nodes share a head. See supplementary materials S1 for a comparison of results obtained from SUD vs. UD data.

We believe the SUD analysis is most appropriate because SUD reflects the syntactic relationships and analyses that were used in the development of theories of formal crossing constraints (Joshi 1985), including early work investigating dependency treebanks (Nivre 2006).

2.3.3 Statistical Method. To test whether real language trees differ from random baseline trees in the distribution of crossing constraints, we fit mixed-effect Poisson regression models. Poisson regression is appropriate for modeling nonnegative integer-valued data such as the formal crossing properties. The dependent variables in the models are the rate of violations of crossing constraints (gap degree, edge degree, etc.).

Suppose that G_{ij} is the gap degree for i^{th} sentence of the j^{th} language, S_{ij} is the length of i^{th} sentence of the j^{th} language, R_{ij} is a dummy variable that encodes whether the sentence is a real tree (as 1) or a baseline tree (as 0), β_0 is the intercept term, β_1 and β_2 are the slope terms for the main effect of sentence length and real/baseline variable, respectively, β_3 is the interaction term, and $u_{0,j}$ is the random intercept adjustment for j^{th} language. The model to predict gap degree, G_{ij} is:

$$\log G_{ij} = (\beta_0 + u_{0,j}) + \beta_1 S_{ij} + \beta_2 R_{ij} + \beta_3 S_{ij} R_{ij} + \epsilon \tag{2}$$

The above model predicts gap degree as a function of sentence length in real and random baselines trees. A similar model is fit for other crossing constraints (well-nestedness, edge degree, etc.). We also vary the predictor variable in place of sentence length, such that a crossing constraint could be a function of sentence length, tree height, or tree arity.

To evaluate evidence for the hypothesis whether gap degree is lower in real trees compared with baseline trees, we compare the model in Equation (2) with a null model that lacks the R_{ij} term, using a likelihood ratio test. We report log-likelihood ratio values, interpreting them as strength of evidence for a difference between real and random trees.

The log-likelihood ratios can be interpreted as logarithmic Bayes factors comparing two hypotheses with equal prior probability: H_0, that there is no distinction between real and random trees; and H_1, that there is a distinction as given by the regression coefficients β_2 and β_3. A higher log-likelihood ratio indicates stronger evidence for H_1.

2.4 Results

The distributions of the formal measures in real trees and random baselines are shown in Figure 7. Table 2 summarizes results in terms of log-likelihood ratios. We find that there is uniformly strong evidence that edge degree, end-point crossings, and HDD are different between real and random trees. This means that the distribution of these formal properties cannot be explained solely in terms of generic constraints on number of crossings, tree topology, and dependency length.

However, for gap degree and well-nestedness, the picture is different. When comparing real trees against DL-controlled random linear arrangements, we do not find substantial evidence for differences in well-nestedness or gap degree. When comparing against random linear arrangements, we do not find substantial evidence for differences in gap degree. These log-likelihood ratios are dramatically smaller than those for, for example, HDD. If there is a distinction in gap degree between real and random trees, the evidence for this distinction is dramatically small compared with other constraints.

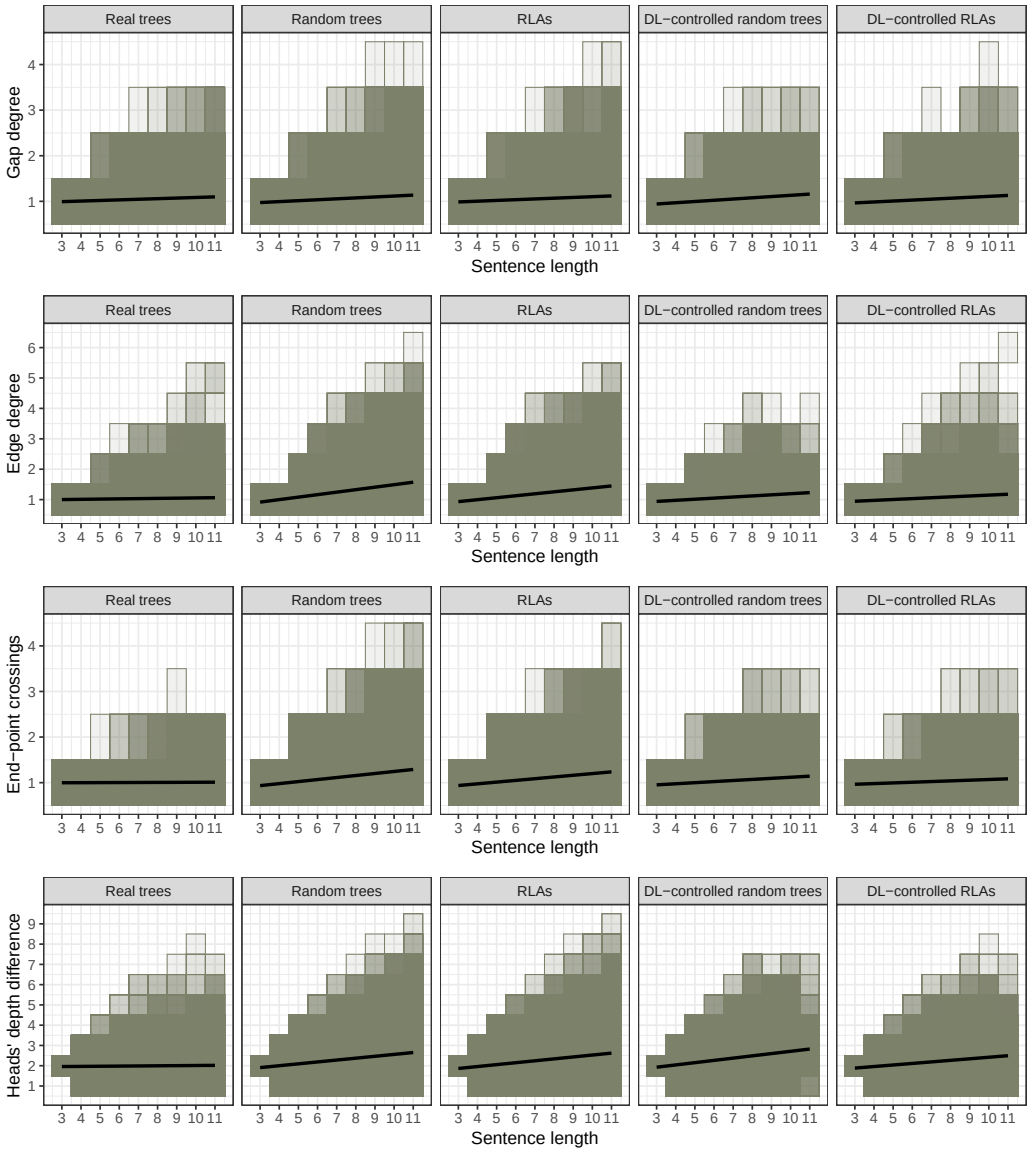


Figure 7
Distribution of graph-theoretical measures gap degree, edge degree, endpoint crossings, and HDD in real trees and several random baselines' trees. Projective trees are excluded from this figure, but included in all regressions. The gradient from light to darker color represent the distribution of a particular crossing constraint for given sentence length. Consider gap degree plot for real trees. The number of trees with gap degree 4 for sentence length 11 is larger than that for sentence length 10, represented by darker shade.

2.5 Discussion

We tested whether there is evidence for graph-theoretic constraints on crossing dependencies beyond what can be explained in terms of other, generic properties of natural language trees.

Table 2
Graph-theoretic constraints on crossing dependencies. The table summarizes the degree of evidence for graph theoretic constraints: To what extent several baselines controlling for the rate of crossing dependencies provide support for graph theoretic constraints on crossing dependencies. Here, the degree of evidence refers to the log-likelihood ratio of the model shown in 2 against a null model that lacks the R_{ij} term.

Baseline	Evidence for constraint on				
	Gap degree	Well-nestedness	Edge degree	End-point crossings	HDD
Random trees					
Sentence length	4	24	740	240	663
Maximum arity	251	45	1,142	465	1,021
Tree height	66	28	296	37	139
Random linear arrangements					
Sentence length	1	8	446	161	590
Maximum arity	1	3	390	133	513
Tree height	3	5	431	167	642
DL-controlled random trees					
Sentence length	0.16	16	17	13	327
Maximum arity	43	21	83	75	498
Tree height	252	58	140	148	49
DL-controlled random linear arrangements					
Sentence length	0.01	2	11	7	181
Maximum arity	0.04	1	7	4	152
Tree height	0.02	2	9	5	164

We find two key results:

1. There is decisive evidence that edge degree, end-point crossings, and HDD are different between real and random trees, suggesting that any constraints on these quantities cannot be explained merely in terms of the number of crossings, tree topology, and dependency length distribution in natural language trees.
2. We find insubstantial evidence for gap degree and well-nestedness constraint after controlling for the number of crossings, tree topology, and dependency length.

Taken together, these results suggest that, despite the massive literature on gap degree and well-nestedness and their connections to formal language theory, these constraints are not the formal properties that most strongly characterize crossing dependencies. In comparison, edge degree, end-point crossings, and HDD—measures that have emerged from the parsing literature, not the formal language theory literature—emerge as strongly characteristic of crossing dependencies. We additionally note an important caveat that our results hold only in short trees, and might be different in larger trees (cf. Ferrer-i-Cancho et al. 2021). The patterns for longer sentences can only be confirmed by baseline generation, which will be taken up in future work.

Among the crossing constraints investigated in the current work, the evidence for HDD as a crossing constraint is strongest. Recall that the HDD constraint is motivated by the findings in the psycholinguistic literature that increased embedding leads to processing difficulty. This suggests that constraints on crossing dependencies could be driven by processing considerations. We turn next to a preliminary investigation where this possibility is explored further.

3. Local Psycholinguistic Factors

In the previous section we investigated the role of formal graph-theoretic factors in determining the distribution of crossing dependencies in natural language. As stated earlier, these factors can be construed as capturing the competence-based constraints on grammar. However, natural language grammar can, in principle, also be influenced by processing-based constraints. One such early proposal can be found in Joshi (1985), where efficient (asymptotic) parsing complexity is a key design requirement for natural languages. More recent proposals include the role of efficient parsing in shaping word order in natural language (see, e.g., Hawkins 2004, although his notion of "efficiency" is quite different from Joshi's). On this account, rarity of crossing dependencies across languages could be assumed to reflect processing difficulty in handling such configurations (also see Ferrer-i-Cancho 2014).

Given the incremental nature of language production (e.g., Ferreira and Henderson 1998), it is reasonable to assume that production of crossing dependencies, which necessarily involves a discontinuity in a phrasal boundary, would incur an increased processing cost compared with non-crossing dependencies. Indeed, recent work investigating filler–gap dependencies (Momma 2021) suggests that such crossing dependencies might require additional cognitive resources in planning. Given the evidence for the tight link between production and comprehension difficulty (MacDonald 2013; Scontras, Badecker, and Fedorenko 2017), it is therefore not far-fetched to assume that comprehenders should also find crossing dependencies difficult to process (Levy et al. 2012; Yadav, Vaidya, and Husain 2017; Husain and Vasishth 2015; Staub et al. 2018). Indeed, it is well known that, given an unbounded dependency, the comprehension system tries to resolve it as soon as possible, a constraint known as active-filler strategy (Frazier 1987). The experiments discussed in this section, therefore, explore whether processing factors such as working-memory constraints, predictability, and so on, could play a role in determining the occurrence of a crossing dependency in the input.

3.1 Motivation

The key motivation of this preliminary study is to investigate if processing related factors modulate the occurrence of a crossing dependency. We investigate two factors—namely, working-memory constraints and prediction processes.

The influence of working-memory limitations on sentence comprehension as well as production is well attested. Cross-linguistically, the linear distance between syntactically related words (i.e., dependency length) has been found to be minimized, in the phenomenon of dependency length minimization (for reviews, see Liu, Xu, and Liang 2017; Temperley and Gildea 2018). This dependency length minimization has recently been argued to be a manifestation of information locality (Futrell 2019; Hahn, Degen, and Futrell 2021), the key idea being that words with high pointwise mutual information (PMI) tend to be close to each other. The "closeness" between a pair of words is usually operationalized as linear distance between them, but more generally

"closeness" would mean the simplicity of the structure that intervene a pair of words (see Yadav, Mittal, and Husain 2020). This generalized version of information locality can be termed as information-simplicity: The words with high PMI tend to have simpler intervening structure between them.

Together, dependency length minimization and information-simplicity have clear implications for crossing dependencies. First, the **information-simplicity hypothesis** predicts that a head-dependent pair with high PMI is less likely to be involved in a crossing dependency. Second, dependency length minimization implies that long phrases could be extraposed in order to avoid increased dependency distance in situ. We call this heavy-phrase extraposition. The **heavy-phrase extraposition hypothesis** predicts that dependencies with a heavy dependent, in terms of length of the phrase, are more likely to be involved in a crossing dependency. Additionally, we expect a positive correlation between dependency length and crossing tendency: The dependencies which are shorter in length are less likely involved in a crossing configuration (Ferrer-i-Cancho 2014; Ferrer-i-Cancho and Gómez-Rodríguez 2016). We call this **localized-simplicity hypothesis**: Localized words have simpler intervening structures between them.

Apart from the role of working memory, another factor that has garnered much attention in the processing literature is prediction. Sentence processing is known to involve a robust top–down component that involves a preactivation of upcoming linguistic material. Such predictions are known to facilitate comprehension (Levy 2008; Smith and Levy 2013) and to attenuate the cost of memory constraints (Husain, Vasishth, and Srinivasan 2014). On this account, when the presence of a dependent is highly expected, this high expectation could offset any cost incurred due to crossing dependency (Levy et al. 2012). So, we test the **expectation hypothesis** that a crossing configuration is more likely in situations where the upcoming dependent/head is highly expected. Expectation of a dependent given a head is operationalized as the log probability that a head has at least one outgoing dependency with that relation type.

Given the gap between the point at which prediction is made and the point at which the concerned linguistic entity is received via input, linguistic predictions have to be maintained in memory (Gibson 1998). This maintenance cost has been shown to correspond to measurable processing difficulty (Husain, Vasishth, and Srinivasan 2015; Ristic et al. 2021), that is, prediction maintenance over a longer period can be costly. It is therefore expected that longer maintenance of dependencies involved in crossing should be avoided. In this work we operationalize maintenance as (a) the number of words between the heads of the crossing dependencies, and (b) the number of heads between the heads of the crossing dependencies.[3] We expect that dependencies that involve a crossing configuration tend to have shorter distance between their heads both in terms of number of words and number of heads. We call this **head-head locality hypothesis**: Two dependencies with their heads being far away from each other are less likely to form a crossing configuration.

3.2 Methods

In order to assess the role of various processing factors mentioned above, we fit logistic regression models to predict whether two dependencies cross.

3 The tendency to minimize the number of heads in such configuration could be related to a recent finding that syntactic heads are avoided in the intervening regions of a dependency (Yadav, Mittal, and Husain 2020).

(a) C_{ij} **= 0** Dependency j **does not** cross dependency i

(b) C_{ij} **= 1** Dependency j crosses dependency i

Figure 8
Schematic showing dependency i and dependency j in two sentences (a) and (b). In (a), dependency j does not cross dependency i; while in (b), j crosses i. We assume that certain properties of dependency i and dependency j determine whether j crosses i or not. We call dependency i a host dependency and dependency j a crosser dependency.

Given dependency i and dependency j in a sentence (see Figure 8), this model uses the following predictors to predict whether dependency i is crossed by dependency j or not. We call i a host dependency and j a crosser dependency.

1. Head-dependent pointwise mutual information of dependency i and dependency j, i.e., PMI_i and PMI_j
2. Expectation of seeing a dependency relation given head type for dependency i and dependency j, i.e., EXP_i and EXP_j
3. Distance between head and dependent of dependency i and dependency j, i.e., DD_i and DD_j
4. Weight of dependent of dependency i, W_i, i.e., the number of nodes transitively dominated by the dependent of i plus one
5. Linear distance between head of dependency i and head of dependency j, LHD
6. Hierarchical distance (number of heads) between head of dependency i and head of dependency j, HHD

Given the variable mentioned above, suppose that C_{ij} indicates whether dependency i is crossed by dependency j of a sentence or not. The model to predict C_{ij} is:

$$\begin{aligned} logit(C_{ij}) = \beta_0 \; + \; & \beta_1 PMI_i + \beta_2 EXP_i + \beta_3 DD_i + \beta_4 W_i + \beta_5 PMI_j + \beta_6 EXP_j \\ & + \beta_7 DD_j + \beta_8 LHD + \beta_9 HHD \end{aligned} \tag{3}$$

Given the discussion in Section 3.1, we predict the following:

- The information-simplicity hypothesis predicts that a dependency with high pointwise mutual information is less likely to get involved in a crossing construction. Thus, it predicts a negative estimate for the parameters β_1 and β_5.
- The heavy-phrase extraposition hypothesis predicts that if a dependency has a heavy dependent, it is more likely to get crossed. Thus, a positive estimate is predicted for β_4.

- The localized-simplicity hypothesis predicts a positive correlation between dependency length and crossing tendency, that is, a positive estimate for β_3 and β_7.
- The expectation hypothesis predicts that a dependency with high expectation is more likely to involve a crossing construction, namely, a positive estimate for β_2 and β_6.
- Finally, the head-head locality hypothesis predicts that two dependencies with longer distance between their heads (i.e., higher LHD or HHD) are less likely to cross each other, that is, β_8 and β_9 should show a negative estimate.

3.2.1 Data. We choose 12 languages from the Surface-syntactic Universal Dependencies (SUD v2.4) treebank, out of which 7 are head-initial and 5 are head-final languages. The criteria for language selection is based on corpus size (>5,000 trees) and language typology. Out of the 7 languages, Hindi, Dutch, Japanese, German, and Korean are head-final languages; the remaining languages, English, Arabic, French, Spanish, Italian, Polish, and Romanian, are head-initial.

3.3 Results

Table 3 summarizes results from the logistic regression model. Results show that all the predictors, namely, pointwise mutual information, expectation, dependency length, dependent weight, distance between the head of the host, and head of the crosser dependency, have significant effect on crossings in the expected directions.

We find an effect of pointwise mutual information for most of the languages such that two dependencies tend to cross each other if their head-dependent mutual information is low. In other words, a dependency that has relatively low mutual information

Table 3
Effect of predictability, locality, inter-head distance on the tendency of two dependencies to cross. The table shows logistic regression coefficients. The significant effects are shown in bold. PMI stands for pointwise mutual information between the head and the dependent in a dependency. DD stands for dependency distance.

Language	Host dependency				Crosser dependency			LHD	HHD
	PMI	EXP	DD	dep.Weight	PMI	EXP	DD		
English	**−0.26**	**−0.11**	**0.09**	**0.014**	**−0.27**	**−0.09**	**0.09**	−0.007	**−0.147**
Hindi	**−0.03**	**−0.31**	**0.07**	**0.015**	**0.15**	**−0.25**	**0.08**	**0.029**	**−0.192**
Dutch	**−0.08**	**−0.09**	**0.10**	**0.020**	**0.04**	**−0.12**	**0.11**	**−0.091**	**0.085**
Arabic	**−0.23**	**0.02**	**0.06**	**0.045**	**0.17**	**−0.23**	**0.08**	**−0.027**	**−0.103**
Japanese	**−0.40**	**0.02**	**0.05**	**0.031**	**−0.39**	**−0.08**	**0.04**	**−0.023**	**−0.151**
French	**−0.06**	**−0.30**	**0.08**	**0.018**	**−0.05**	**−0.34**	**0.08**	**0.068**	**−0.278**
German	**−0.20**	**−0.17**	**0.11**	**0.031**	**−0.03**	**−0.16**	**0.11**	**−0.069**	**0.061**
Korean	**−0.05**	**−0.07**	**0.12**	**0.082**	**−0.48**	0.01	**0.09**	**−0.04**	**−0.45**
Spanish	**−0.11**	**−0.11**	**0.08**	**0.012**	**−0.19**	**−0.01**	**0.09**	**0.01**	**−0.21**
Italian	**0.05**	**−0.22**	**0.09**	**0.021**	**0.08**	**−0.23**	**0.10**	**−0.02**	**−0.09**
Polish	**−0.09**	**−0.11**	**0.11**	**0.010**	**−0.05**	**−0.09**	**0.12**	0.002	**0.01**
Romanian	**−0.08**	**−0.09**	**0.12**	**0.028**	**−0.05**	**−0.09**	**0.14**	**−0.02**	**−0.06**

between its head and dependent is more likely to cross—and get crossed by—another dependency. The result supports the information-simplicity hypothesis.

Additionally, all 12 languages show an effect of the dependent weight such that a dependency with heavier dependent phrase is more likely to get crossed by another dependency. The result supports the heavy-phrase extraposition hypothesis.

The results also support the localized-simplicity hypothesis, that is, dependency distance is positively correlated with crossing tendency—a shorter dependency is less likely to be involved in a crossing construction.

With regard to the role of prediction, results show that a dependency relation that has relatively low expectation given a head type is more likely to cross—and get crossed by—another dependency. The result does not support the expectation hypothesis. Finally, we find that the linear and hierarchical distance between the heads of two dependencies negatively influence whether they cross or not. Crossing tendency reduces if the number of heads/words between the heads of the two dependencies increases. This result holds for most of the languages (except Hindi, Spanish, Dutch, and Polish). The LHD/HHD results support the head-head locality hypothesis motivated by prediction maintenance account.

3.4 Extraposition in Noun Phrases

We also test a prediction by Levy et al. (2012) that a crossing dependency is easier to comprehend if expectation of a dependency relation given a nominal head is higher. We test whether expectation has a positive effect on crossing tendency of noun-headed constructions (see Figure 9). In addition to expectation effect, we also test the effect of weight of dependent of noun phrase—a noun-headed dependency is more likely to get crossed if its dependency is heavier because the heavy dependent may move around to minimize dependency length. To test the hypothesis, we fit a logistic regression similar to 3 with two predictors, namely, expectation of dependency relation and dependent weight.

Table 4 shows the estimates of effect of expectation and dependent weight on crossing tendency of a noun-headed construction. The effect of expectation has a negative estimate (except Japanese), suggesting that if expectation of seeing a particular dependency relation in a noun-headed dependency is higher, the dependency is less likely to get crossed by another dependency.

The weight of dependent in a noun-headed construction positively affects its crossing tendency (except Japanese)—a noun-headed construction has a higher tendency to get crossed by another dependency if the weight of its dependent is higher. The result supports the heavy-phrase extraposition hypothesis: A heavy dependent is more

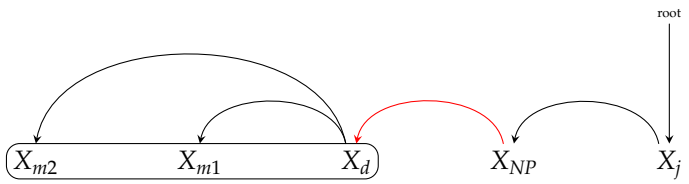


Figure 9
A noun-headed construction: the node X_{NP} is a noun phrase and is modified by a dependent X_d. The dependent X_d has three nodes in its projection, namely, X_{m2}, X_{m1}, and X_d, so the weight of the dependent X_d is 3.

Table 4
Effect of expectation and host dependent weight on the tendency of an noun-headed construction to be a crossing.

Language	Dependent weight			Expectation			D.weight x Expectation		
	$\hat{\beta}_1$	SE	Z value	$\hat{\beta}_2$	SE	Z value	$\hat{\beta}_3$	SE	Z value
English	0.096	0.013	7.36 *	−0.431	0.018	−22.81 *	0.005	0.003	2.04 *
Hindi	0.103	0.005	20.05 *	−0.422	0.011	−39.09 *	0.013	0.001	9.61 *
Dutch	0.122	0.006	18.89 *	−0.444	0.014	−31.41 *	0.005	0.001	3.77 *
Arabic	0.064	0.006	10.78 *	−0.250	0.025	−9.89 *	0.004	0.001	2.85 *
Japanese	0.046	0.002	22.57 *	−0.006	0.011	−0.61 n.s.	−0.001	0.001	−1.22 n.s.
French	0.110	0.004	22.61 *	−0.486	0.011	−41.77 *	0.008	0.001	8.96 *
German	0.198	0.005	36.29 *	−0.499	0.016	−29.86 *	0.016	0.001	13.93 *
Korean	0.009	0.012	0.82 n.s.	0.054	0.017	3.11 *	−0.029	0.003	−9.70 *
Spanish	0.102	0.005	19.39 *	−0.067	0.026	−2.51 *	0.005	0.002	2.59 *
Italian	0.092	0.006	14.34 *	−0.400	0.016	−24.17 *	0.003	0.001	2.60 *
Polish	−0.016	0.011	−1.39 n.s.	−0.051	0.016	−3.10 *	−0.008	0.002	−4.67 *
Romanian	0.143	0.011	12.07 *	−0.212	0.034	−6.17 *	0.018	0.005	3.76 *

likely to get extraposed—hence causing a crossing dependency—in order to minimize dependency distance.[4]

We also find a significant interaction effect between dependent weight and expectation such that effect of dependent weight increases with increase in expectation. This result suggests that high expectation facilitates crossing dependencies with heavy dependent. In other words, a heavy dependent is more likely to get extraposed when expectation is high. This pattern has some support for the hypothesis proposed by Levy et al. (2012) that increased expectation facilitates the processing of right-extraposed structures.

3.5 Comparison with Baseline Trees

One concern regarding the significant effect of formal measures like dependency distance, dependent weight, and inter-head distance could be that they do not capture any linguistic or cognitive phenomenon, rather, their effect could be due to independent reasons such as tree topology. For example, it is possible that an arc in any random directed acyclic graph is more likely to get crossed by another arc if linear distance between its nodes is larger (Ferrer-i-Cancho and Gómez-Rodríguez 2016).

To check whether some natural language phenomena underlies the effects of dependency distance, dependent weight, and so forth, on crossing dependency, we fit a model to predict crossing tendency in real vs. random trees. The model includes interaction effect terms to test whether the expected effect of dependency distance, dependent weight, and so on, is larger in real trees compared with the random trees generated as part of Section 2. If the effect of these predictors arise due to some real psycholinguistic phenomena, we expect a positive estimate for the interaction effect.

Table 5 shows the results of the model. We find a significant interaction effect for all the predictors.

4 One can argue that effect of dependent weight in noun-headed constructions is only due to right-extraposed relative clauses and hence the result should not be generalized to heavy-phrase extraposition hypothesis. However, the results for noun-headed constructions remains the same when we remove relative clause cases—implying that weight-expectation effect in noun-headed constructions is not driven by relative clauses constructions alone.

Table 5
Effect of dependency distance, dependent weight, and inter-head distance on crossing tendency in real trees compared with random trees.

Predictor	Estimate	Std. Error	z value	
Effect of predictors in random trees				
DD (host dependency)	0.27	0.003	89.91	*
DD (crosser dependency)	0.21	0.003	69.45	*
LHD	−0.06	0.006	−9.82	*
HHD	0.05	0.014	3.53	*
Host dependent weight	−0.28	0.005	−51.03	*
Interaction effects: Effect of predictors in real trees compared with random trees				
DD: Real (host dependency)	0.15	0.015	10.17	*
DD: Real (crosser dependency)	0.18	0.015	11.77	*
LHD: Real	0.35	0.025	13.84	*
HHD: Real	−0.30	0.052	−5.78	*
Host dependent weight: Real	0.31	0.022	13.75	*

- The effect of dependency distance on crossing tendency is larger in real language trees, such that longer dependencies are more likely to cross or get crossed by another dependency in real trees compared to random trees.
- The effect of dependent weight is driven by real trees, such that a dependency with heavy dependent is more likely to get crossed in real trees, but not in random trees (the main effect is in opposite direction).

3.6 Discussion

The results from the logistic regression models provide compelling evidence for the role of working-memory constraints and information-locality on the occurrence crossing dependencies. To summarize, we find support for:

1. **Information simplicity:** The words with high mutual information are less likely to form crossing configurations.
2. **Localized simplicity:** The words that are close to each other are less likely to involve a crossing dependency.
3. **Heavy-phrase extraposition:** A dependency with a heavy dependent is more likely to get crossed by another dependency.
4. **Head-head locality:** Two dependencies with their heads being far away from each other are less likely to form a crossing configuration.

However, the results were only partly consistent with the expectation hypothesis: High expectation for the presence of a dependent correlates with a higher rate of crossing only when that dependent is heavy, as indicated by the presence of a positive interaction of weight and expectation in Table 5 for most languages, but a negative main effect of expectation. Overall, we found evidence for working-memory based and information-locality based accounts, but limited evidence for expectation-based accounts.

An interesting result that should be interpreted in the light of previous findings is the effect of dependency lengths on the occurrence of crossing dependencies. Recent work has shown that a constraint on dependency length alone cannot fully explain the low rate of crossing dependencies in natural languages (Yadav, Husain, and Futrell 2021). In contrast, our results show an effect of dependency length on crossing tendency such that longer dependencies are more likely to cross—or get crossed by—other dependencies. Additionally, our results in Section 2 show that dependency lengths can explain ill-nestedness distribution in natural languages. Together, these findings suggest that although dependency length affects the occurrence of crossing dependencies, it cannot fully explain the observed quantitative distribution of crossings in natural languages.

4. General Discussion

The current work investigated crossing dependencies from two perspectives. These were (a) the role of certain global graph-theoretic factors in determining the distribution of crossing dependencies in natural language, and (b) the role of certain local psycholinguistic factors in determining if two dependencies cross. Results from Experiment 1 provide strong evidence that edge-degree, end-point crossings, and HDD determine the distribution of crossing dependencies in natural language, while they provide weak to insubstantial evidence that other factors such as gap-degree and well-nestedness constrain crossing dependencies. This suggests that the apparent bounds on gap-degree and well-nestedness arise as a consequence of factors such as number of crossings, tree topology, and dependency length.

In particular, the results from Experiment 1 provide the strongest evidence in favor of cognitively motivated crossing constraints (such as HDD), suggesting a parsing/processing driven constraint on crossing dependencies. This proposal was further substantiated through the results from Experiment 2 that show that certain psycholinguistically motivated factors such as information locality, dependency weight, as well as inter-head distance can determine if a pair of dependency will cross, even when compared against random trees. These results provide evidence for a functional motivation for distribution of crossing dependencies in natural language. The key contribution of this work is to (a) provide a method to quantify the evidence for a particular graph theoretic constraint beyond what can be explained in terms of tree topology, number of crossings, and dependency distance, and (b) to highlight that crossings can be predicted through processing factors beyond dependency distance.

Together, the two findings suggest that graph-theoretic constraints on crossing dependencies could be driven by processing considerations (also see Ferrer-i-Cancho and Gómez-Rodríguez 2016). The idea that formal properties in grammar could be determined by processing consideration is not new. As stated earlier, Joshi (1985) proposed the MCS hypothesis, which required grammars to be efficiently parseable, in the sense of worst-case asymptotic complexity of exact parsing. Similarly, the performance-grammar correspondence hypothesis by Hawkins (2004) proposes that processing strategies get grammaticalized for efficiency considerations. We note that, on these accounts, efficiency in grammar is understood as efficiency in online comprehension of the utterances licensed by the grammar.

There is evidence that humans find crossing dependencies to be difficult during comprehension. One piece of evidence for this comes from processing of filler–gap dependencies. A filler-gap dependency typically involves a crossing dependency and, in order to resolve it, the human processing system is known to operate with the principle of immediacy, namely, the parser tries to resolve it sooner rather than later

(De Vincenzi 1991). There is also evidence that, if possible, such dependencies are avoided (Staub et al. 2018). More recently, Husain and Yadav (2020) show that in Hindi, crossing dependencies are avoided during comprehension of participle clauses and that such dependencies lead to processing difficulty. Processing of crossing dependency necessarily involves maintenance of unresolved structure and a processing strategy to avoid building such complex configuration is consistent with the parser's bias for building simple structures (Frazier 1985). In the domain of computational parsers, limits on quantities such as edge-degree, end-point crossings, and so forth, have been previously shown to lead to efficiency in terms of both asymptotic complexity and practical accuracy (Pitler, Kannan, and Marcus 2013), but there has not yet been any systematic investigation on whether these factors affect human online processing. If it is true that constraints on crossing dependencies are motivated by processing efficiency for humans, then formal factors such as end-point crossings might also lead to observable processing difficulty for humans.

The research discussed above suggests that the distribution of crossing dependencies may arise due to comprehension difficulty, but processing efficiency is not limited to comprehension only: Production and learning may also play a role. For example, it is known that difficulty during comprehension could be a consequence of pressures during production (MacDonald 2013); in this view, ease of production determines the distribution of linguistic patterns in a language community, and this distribution in turn makes certain structures either easy or difficult to comprehend. Learning may also play a role: Given that crossing dependencies are difficult to produce and comprehend, it is possible that such dependencies are also difficult to learn and that their rarity could be independently driven by learning biases (Chang 2009). Teasing apart the role (and extent) of how various processing factors affect crossing dependencies will require a dedicated research effort. We hope that our work is a step in that direction.

If crossing dependencies are difficult to process, then why do they exist at all? To answer this question, we need to appreciate that processing cost due to crossings could be just one of the many sources that lead to processing difficulty. Decades of research on comprehension and production has shown that processing complexity can be understood as a trade-off between various countervailing factors related to sentence encoding/decoding (e.g., Trueswell, Tanenhaus, and Garnsey 1994; Kaiser and Trueswell 2004; Altmann and Kamide 1999; Frazier 1979; Gibson 1998; Lewis and Vasishth 2005; Levy 2008). For example, syntactic configurations leading to clausal embeddings are known to be quite complex and under such configurations creating a crossing dependency (for example, by right-extraposition) could in fact lead to a *relatively* less complex structure (Yngve 1960). This, of course, implies that in certain context, other syntactic configurations could be costlier. Similarly, requirements at different levels of linguistic encoding (discourse, pragmatics, etc.) could require positioning of words that could lead to creation of crossing dependencies. Finally, during production, accessibility-related pressures (e.g., Ferreira and Dell 2000; Branigan, Pickering, and Tanaka 2008) could lead to creation of crossing dependencies. In short, formation of crossing dependencies could be due to (a) creation of less costly structures, (b) extra-sentential requirements, and (c) production pressures such as accessibility.

Could crossing dependencies arise due to non-functional reasons? It is possible that a language-wide grammatical constraint (e.g., on the position of heads) could lead to existence of certain crossing dependencies. For example, in English, the presence of *wh* crossing dependencies could be a consequence of a fixed word order in the language. Compare this to *wh* dependencies in a free word order language like Hindi where the *wh* phrase can appear in situ, thereby creating a non-crossing dependency. For such

cases, avoiding a crossing dependency in a language like English is not possible due to language-wide grammatical constraints.

We hope this work stimulates future work studying the functional bases for restrictions on crossing dependencies using experimental and corpus methods.

Acknowledgments

We thank the three anonymous reviewers for helpful suggestions. This work was supported by an NVIDIA GPU grant to RF.

References

Ades, A. E. and M. J. Steedman. 1982. On the order of words. *Linguistics and Philosophy*, 4:517–558. https://doi.org/10.1007/BF00360804

Altmann, G. and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Bach, Emmon, Colin Brown, and William Marslen-Wilson. 1986. Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 1:249–262. https://doi.org/10.1080/01690968608404677

Bodirsky, Manuel, Marco Kuhlmann, and Mathias Möhl. 2005. Well-nested drawings as models of syntactic structure. In *Tenth Conference on Formal Grammar and Ninth Meeting on Mathematics of Language*, pages 195–203.

Boston, Marisa Ferrara, John T. Hale, and Marco Kuhlmann. 2010. Dependency structures derived from Minimalist grammars. In *Proceedings of the 10th and 11th Biennial Conference on The Mathematics of Language*, pages 1–12. https://doi.org/10.1007/978-3-642-14322-9_1

Branigan, Holly P., Martin J. Pickering, and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2):172–189. https://doi.org/10.1016/j.lingua.2007.02.003

Bresnan, Joan. 1982. *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.

Bresnan, J., R. M. Kaplan, S. Peters, and A. Zaenen. 1982. Cross-serial dependencies in Dutch. *Linguistic Inquiry*, 13:613–636. https://doi.org/10.1007/978-94-009-3401-6_11

Chang, Franklin. 2009. Learning to order words: A connectionist model of Heavy NP Shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61:374–397. https://doi.org/10.1016/j.jml.2009.07.006

Chomsky, Noam. 1956. Three models for the description of language. *Information Theory, IRE Transactions On*, 2(3):113–124. https://doi.org/10.1109/TIT.1956.1056813

Chomsky, Noam. 1957. *Syntactic Structures*. Walter de Gruyter. https://doi.org/10.1515/9783112316009

Chomsky, N. 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht, The Netherlands.

Chomsky, Noam. 1995. *The Minimalist Program*, volume 28. Cambridge University Press.

Chomsky, Noam. 2007. Approaching UG from below. In *Interfaces + Recursion = Language?: Chomsky's Minimalism and the View from Syntax-Semantics*. Mouton de Gruyter, pages 1–29. https://doi.org/10.1515/9783110207552-001

Chomsky, Noam and Marcel P. Schützenberger. 1963. The algebraic theory of context free languages. In P. Braffot and D. Hirschberg, editors, *Computer Programming and Formal Languages*. North Holland, Amsterdam, pages 118–161. https://doi.org/10.1016/S0049-237X(08)72023-8

De Vincenzi, M. 1991. *Syntactic Parsing Strategies in Italian*. Kluwer Academic, Dordrecht, The Netherlands. https://doi.org/10.1007/978-94-011-3184-1

Eisner, Jason and Giorgio Satta. 1999. Efficient parsing for bilexical context-free grammars and head automaton grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 457–464. https://doi.org/10.3115/1034678.1034748

Ferreira, Fernanda and John M. Henderson. 1998. Linearization strategies during language production. *Memory & Cognition*, 26(1):88–96. https://doi.org/10.3758/BF03211372

Ferreira, V. S. and G. S. Dell. 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4):296–340. https://doi.org/10.1006/cogp.1999.0730

Ferrer-i-Cancho, Ramon. 2006. Why do syntactic links not cross? *Europhysics Letters (EPL)*, 76(6):1228–1235. https://doi.org/10.1209/epl/i2006-10406-0

Ferrer-i-Cancho, Ramon. 2014. Non-crossing dependencies: Least effort, not grammar. *CoRR*, abs/1411.2645.

Ferrer-i-Cancho, Ramon. 2014. A stronger null hypothesis for crossing dependencies. *EPL (Europhysics Letters)*, 108(5):58003. https://doi.org/10.1209/0295-5075/108/58003

Ferrer-i-Cancho, Ramon, Christian Bentz, and Caio Seguin. 2020. Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, pages 1–30. https://doi.org/10.1080/09296174.2020.1778387

Ferrer-i-Cancho, Ramon, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2018. Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, 493:311–329. https://doi.org/10.1016/j.physa.2017.10.048

Ferrer-i-Cancho, Ramon and Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328. https://doi.org/10.1002/cplx.21810

Ferrer-i-Cancho, Ramon, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. 2022. The optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308. https://doi.org/10.1103/PhysRevE.105.014308

Ferrer-i-Cancho, Ramon, Antoni Hernández-Fernández, David Lusseau, Govindasamy Agoramoorthy, Minna J. Hsu, and Stuart Semple. 2013. Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578. https://doi.org/10.1111/cogs.12061

Frazier, Lynn. 1979. On comprehending sentences: Syntactic parsing strategies. *ETD Collection for University of Connecticut*.

Frazier, Lyn. 1985. Syntactic complexity. *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 129–189. https://doi.org/10.1017/CBO9780511597855.005

Frazier, Lyn. 1987. Syntactic processing: Evidence from Dutch. *Natural Language & Linguistic Theory*, 5(4):519–559. https://doi.org/10.1007/BF00138988

Futrell, Richard. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15. https://doi.org/10.18653/v1/W19-7902

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74. https://doi.org/10.18653/v1/W18-6008

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): Surface-syntactic relations and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks & Linguistic Theory*, pages 126–132. https://doi.org/10.18653/v1/W19-7814

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, Edward, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger P. Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gibson, Edward and James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248. https://doi.org/10.1080/016909699386293

Gómez-Rodríguez, Carlos, John Carroll, and David Weir. 2011. Dependency parsing schemata and mildly non-projective dependency parsing. *Computational Linguistics*, 37(3):541–586. https://doi.org/10.1162/COLI_a_00060

Gómez-Rodríguez, Carlos, Morten H. Christiansen, and Ramon Ferrer-i-Cancho. 2019. Memory limitations are hidden in grammar. *CoRR*, abs/1908.06629.

Gómez-Rodríguez, Carlos and Ramon Ferrer-i-Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96:062304. https://doi.org/10.1103/PhysRevE.96.062304

Gómez-Rodríguez, Carlos, Tianze Shi, and Lillian Lee. 2018. Global transition-based non-projective dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 2664–2675. https://doi.org/10.18653/v1/P18-1248

Groß, Thomas and Timothy Osborne. 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, 22:43–90.

Hahn, Michael, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 128(4):726–756. https://doi.org/10.1037/rev0000269

Havelka, Jiří. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 608–615.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hays, David G. 1964. Dependency theory: A formalism and some observations. *Language*, 40:511–525. https://doi.org/10.2307/411934

Hofmeister, Philip and Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language*, 86:366–415. https://doi.org/10.1353/lan.0.0223

Holan, Tomas, Vladislav Kubon, Karel Oliva, and Martin Plátek. 1998. Two useful measures of word order complexity. In *Processing of Dependency-Based Grammars*, pages 21–28.

Hopcroft, John E. and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley.

Hotz, Günter and Gisela Pitsch. 1996. On parsing coupled-context-free languages. *Theoretical Computer Science*, 161(1-2):205–233. https://doi.org/10.1016/0304-3975(95)00114-X

Hudson, Richard A. 1990. *English Word Grammar*. Blackwell.

Husain, Samar and Shravan Vasishth. 2015. Non-projectivity and processing constraints: Insights from Hindi. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 141–150.

Husain, Samar, Shravan Vasishth, and Narayanan Srinivasan. 2014. Strong expectations cancel locality effects: Evidence from Hindi. *PLOS ONE*, 9(7):e100986. https://doi.org/10.1371/journal.pone.0100986

Husain, Samar, Shravan Vasishth, and Narayanan Srinivasan. 2015. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, 8(2):1–12. https://doi.org/10.16910/jemr.8.2.3

Husain, Samar and Himanshu Yadav. 2020. Target complexity modulates syntactic priming during comprehension. *Frontiers in Psychology*, 11:454. https://doi.org/10.3389/fpsyg.2020.00454

Jaeger, T. Florian and Harry J. Tily. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335. https://doi.org/10.1002/wcs.126

Joshi, Aravind K. 1985. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge University Press, Cambridge, pages 190–205. https://doi.org/10.1017/CBO9780511597855.007

Joshi, Aravind K. 1990. Processing crossed and nested dependencies: An automaton perspective on the psycholinguistic results. *Language and Cognitive Processes*, 5:1–27. https://doi.org/10.1080/01690969008402095

Joshi, Aravind K., K. Vijay Shanker, and David Weir. 1991. The convergence of mildly context-sensitive grammar formalisms. In P. Sells, S. Shieber, and Thomas Wasow, editors, *Foundational Issues in Natural Language Processing*. MIT Press, Cambridge, MA, pages 31–81.

Kaiser, E. and J. C. Trueswell. 2004. The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2):113–147. https://doi.org/10.1016/j.cognition.2004.01.002

Kuhlmann, Marco. 2007. *Dependency Structures and Lexicalized Grammars*. Ph.D. thesis, Universität des Saarlandes.

Kuhlmann, Marco. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387. https://doi.org/10.1162/COLI_a_00125

Kuhlmann, Marco and Joakim Nivre. 2006. Mildly non-projective dependency

structures. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 507–514. https://doi.org/10.3115/1273073.1273139

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, Roger P., Evelina Fedorenko, Mara Breen, and Edward Gibson. 2012. The processing of extraposed structures in English. *Cognition*, 122(1):12–36. https://doi.org/10.1016/j.cognition.2011.07.012

Lewis, Richard L. and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419. https://doi.org/10.1207/s15516709cog0000_25

Liu, Haitao, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193. https://doi.org/10.1016/j.plrev.2017.03.002

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4:226. https://doi.org/10.3389/fpsyg.2013.00226

Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126:313–318. https://doi.org/10.1016/j.cognition.2012.09.010

Marcus, Solomon. 1965. Sur la notion de projectivité. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 11(2):181–192. https://doi.org/10.1002/malq.19650110212

Mel'čuk, Igor' Aleksandrovič. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press.

Michaelis, Jens. 1998. Derivational Minimalism is mildly context-sensitive. In *Logical Aspects of Computational Linguistics*, volume 98, pages 179–198, Springer. https://doi.org/10.1007/3-540-45738-0_11

Momma, Shota 2021. Filling the gap in gap-filling: Long-distance dependency formation in sentence production. *Cognitive Psychology*, 129:101411. https://doi.org/10.1016/j.cogpsych.2021.101411

Nivre, Joakim. 2006. Constraints on non-projective dependency parsing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–80.

Nivre, Joakim. 2007. Incremental non-projective dependency parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 396–403.

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer. https://doi.org/10.1007/978-3-319-18111-0_1

Nivre, Joakim, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. *Universal Dependencies 1.0*. Universal Dependencies Consortium.

Phillips, Colin, Nina Kazanina, and Shani H. Abada. 2005. ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research*, 22(3):407–428. https://doi.org/10.1016/j.cogbrainres.2004.09.012

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pitler, Emily, Sampath Kannan, and Mitchell Marcus. 2013. Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics*, 1:13–24. https://doi.org/10.1162/tacl_a_00206

Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Center for the Study of Language and Information, Stanford, CA.

Prüfer, Heinz. 1918. Neuer Beweis eines Satzes über Permutationen. *Archiv der Mathematischen Physik*, 27:742–744.

Ristic, Bojana, Simona Mancini, Nicola Molinaro, and Adrian Staub. 2021. Maintenance cost in the processing of subject–verb dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000863

Sag, Ivan A., Thomas Wasow, and Emily M. Bender. 1999. *Syntactic Theory: A Formal Introduction*. Center for the Study of Language and Information, Stanford, CA.

Satta, Giorgio. 1992. Recognition of linear context-free rewriting systems. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 89–95. https://doi.org/10.3115/981967.981979

Scontras, Gregory, William Badecker, and Evelina Fedorenko. 2017. Syntactic complexity effects in sentence production: A reply to Macdonald, Montag, and Gennari (2016). *Cognitive Science*, 41(8):2280–2287. https://doi.org/10.1111/cogs.12495

Seki, Hiroyuki, Takashi Matsumara, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229. https://doi.org/10.1016/0304-3975(91)90374-B

Smith, Nathaniel J. and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Stabler, Edward P. 1997. Derivational minimalism. In *Logical Aspects of Computational Linguistics*, pages 68–95. Springer. https://doi.org/10.1007/BFb0052152

Staub, Adrian, Francesca Foppolo, Caterina Donati, and Carlo Cecchetto. 2018. Relative clause avoidance: Evidence for a structural parsing principle. *Journal of Memory and Language*, 98:26–44. https://doi.org/10.1016/j.jml.2017.09.003

Steedman, Mark and Jason Baldridge. 2011. Combinatory categorial grammar. In *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. pages 181–224. https://doi.org/10.1002/9781444395037.ch5

Temperley, David and Dan Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15. https://doi.org/10.1146/annurev-linguistics-011817-045617

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Librairie C. Klincksieck.

Trueswell, John C., Michael K. Tanenhaus, and S. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318. https://doi.org/10.1006/jmla.1994.1014

Weir, David Jeremy. 1988. *Characterizing Mildly Context-Sensitive Grammar Formalisms*. Ph.D. thesis, University of Pennsylvania.

Yadav, Himanshu, Samar Husain, and Richard Futrell. 2019. Are formal restrictions on crossing dependencies epiphenomenal? In *Proceedings of the 18th International Workshop on Treebanks & Linguistic Theory*, pages 2–12. https://doi.org/10.18653/v1/W19-7802

Yadav, Himanshu, Samar Husain, and Richard Futrell. 2021. Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, 7(s3):20190070. https://doi.org/10.1515/lingvan-2019-0070

Yadav, Himanshu, Shubham Mittal, and Samar Husain. 2020. Dependency length minimization hypothesis revisited. In *26th Architectures and Mechanisms for Language Processing Conference*, e12822.

Yadav, Himanshu, Ashwini Vaidya, and Samar Husain. 2017. Understanding constraints on non-projectivity using novel measures. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 276–286. Linköping University Electronic Press, Pisa, Italy.

Yadav, Himanshu, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science*, 44(4):e12822. https://doi.org/10.1111/cogs.12822

Yli-Jyrä, Anssi Mikael. 2003. Multiplanarity—A model for dependency structures in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 189–200.

Yngve, Victor H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Oxford, UK. Addison-Wesley Press.