Are formal restrictions on crossing dependencies epiphenomenal?

Himanshu Yadav[1], **Samar Husain**[2]*, **and Richard Futrell**[3]*
[1] Department of Humanities & Social Sciences, Indian Institute of Technology Kanpur
[2] Department of Humanities & Social Sciences, Indian Institute of Technology Delhi
[3] Department of Language Science, University of California, Irvine
yadavhimanshu059@gmail.com, samar@hss.iitd.ac.in, rfutrell@uci.edu

Abstract

Characterizing the distribution of crossing dependencies in natural language dependency trees is a crucial task for building parsers and understanding the formal properties of human language. A number of formal restrictions on crossing dependencies have been proposed, including bounds on gap degree, edge degree, and end-point crossings. Here we ask whether the empirical distribution of crossing dependencies in dependency treebanks offers evidence for these formal restrictions as true, independent constraints on dependency trees, or whether the distribution can be explained using other, more generic constraints affecting dependency trees. Specifically, we explore the null hypothesis that crossing dependencies are formally unrestricted, but occur at a low rate. We implement the null hypothesis using random trees where crossing dependencies occur at the same rate as in natural language trees, but without any formal restrictions. We find that this baseline generally does not reproduce the same distribution of gap degree, edge degree, end-point-crossing, and heads' depth difference as real trees, suggesting that these formal constraints are a consequence of factors beyond the rate of crossing dependencies alone.

1 Introduction

In dependency grammar formalisms, the syntactic structure of a sentence is encoded in the form of head–dependent relations. For the most part, the dependents of a given head form a contiguous substring of the sentence, i.e., all the nodes occurring between the head and its dependent are (transitively) dominated by the head. Such dependencies have been termed **projective**. In addition to projective dependencies, we also find instances where the dependents of a head are discontinuous. This happens when a node in the span of a head and its dependents is not (directly or indirectly) dominated by the head. Such dependencies are known as **crossing** or **non-projective dependencies**. Formally, a dependency $X_h \rightarrow X_d$ is deemed crossing if and only if there is at least one node X_i between X_h and X_d that X_h does not dominate. In Figure 1 the dependency arc from the node X_h to its dependent X_d is crossing because X_i is headed by a node (X_j) which is outside the span of $X_h \rightarrow X_d$. Note that all other arcs in the dependency tree shown in Figure 1 are projective. For example, the arc $X_j \rightarrow X_i$ is a projective arc as X_h is dominated by X_j.

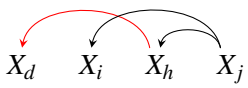

Figure 1: The dependency arc $X_h \rightarrow X_d$ is a crossing dependency. All other arcs are non-crossing.

The most basic cross-linguistic generalization about crossing dependencies is that they are rare (see e.g. Straka et al., 2015). The rarity of crossing dependencies poses several interesting questions that are relevant from formal, computational, and cognitive perspectives. Most fundamentally, why are these

* Equal contribution

constructions rare? When and why are these constructions difficult for computational parsers and humans? Are there general constraints on the space of variation in natural languages that can explain this rarity?

Investigating the constraints which cause the rarity of crossing dependencies could help us in discovering the underlying principles that have shaped human language. Not surprisingly, there have been previous attempts to investigate the cause of this rarity formally as well as from a processing perspective (e.g., Shieber, 1985; Bach et al., 1986; Vogel et al., 1996; Ferrer-i-Cancho, 2006; Levy et al., 2012; Kuhlmann, 2013; Husain and Vasishth, 2015; Ferrer-i-Cancho and Gómez-Rodríguez, 2016; Yadav et al., 2017, under review). In addition, a number of formal restrictions on crossing dependencies have also been proposed. Kuhlmann (2013) proposes that dependency trees have limited *gap degree* and are usually *well-nested* (see Figure 2b). Pitler et al. (2013) propose that crossing dependency configurations have a property called *1-end-point-crossing*. Other 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 et al., 2017). In this paper, we call these formal constraints on crossing dependencies **crossing constraints**.

(a) Projection chains **(b)** Gap degree

Figure 2: The **projection chain** of a node X is the set of all the nodes dominated by X which lie in a single path from X to a terminal node. For example, in the dependency tree **(a)**, $\{X_j,X_h,X_d,X_g\}$ and $\{X_j,X_i,X_k\}$ are two projection chains from the node X_j. A projection chain is continuous if it forms a continuous substring of the sentence. For example, the projection chain of X_h, i.e., $\{X_h,X_d,X_g\}$ is a continuous substring of the sentence $\{X_k,X_i,X_g,X_d,X_h,X_j\}$. The dependency tree **(b)** shows a dependency schema to illustrate gap degree. The gap degree of a node is the largest number of discontinuities in any projection chain. In **(b)**, the projection chain for X_h is $\{X_h,X_d,X_g\}$, which contains 2 discontinuities or gaps, so the gap degree of node X_j is 2.

Crossing constraints are important in two domains: in the development of computational parsers, and for theoretical formal syntax, because these restrictions correspond to the formal language class of natural language. Crossing dependencies indicate deviations from context-free grammar (Marcus, 1965; Shieber, 1985). More specifically, the hierarchy of mildly context-sensitive languages is defined by restrictions on gap degree. Gap degree corresponds to the number of components in a Multiple Context-Free Grammar (Seki et al., 1991) and to the number of distinct selector features in Minimalist Grammars (Michaelis, 1998). It corresponds to the 'limited amount of cross-serial dependencies' allowed in TAG derivations (Joshi, 1985), (also see Bodirsky et al., 2005). In the computational linguistics literature it is common to provide statistics showing that there are only a small number of dependency trees violating any given crossing constraint. For example, Kuhlmann (2013) shows that as gap degree increases, there are fewer and fewer trees per language with that gap degree.

These proposals across the theoretical syntax and parsing literature raise the possibility that crossing constraints might constitute independent, causal constraints on natural language syntax. However, it is also possible that the observed distribution of crossing dependencies may be epiphenomenal, i.e., a consequence of other constraints affecting dependency trees which have nothing to do with crossing dependencies themselves, such as a general pressure to minimize dependency length (e.g., as investigated in Ferrer-i-Cancho and Gómez-Rodríguez, 2016; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). In this paper, we investigate the status of crossing constraints using dependency corpora, asking whether the empirical distribution of crossing dependencies gives evidence for crossing constraints, or whether the data is best explained by an extremely simple null hypothesis: that crossing dependencies are formally unrestricted but simply rare.

As an example of how crossing constraints might be epiphenomenal, consider gap degree. Gap degree refers to the number of discontinuities in the projection chain headed by a node (see Figure 2). So, for example, if the longest projection chain in a sentence is of length 6, then gap degree cannot exceed 5. Now suppose that linguistic dependency trees typically have short projection chains and that crossing dependencies are rare but randomly distributed across dependency trees. Then it is unlikely that we will observe a projection with many discontinuities, simply due to the fact that projection chains are usually short; so we will measure low gap degree. From this measurement, we might falsely conclude that there exists a bound on gap degree. These considerations suggest that gap degree might not have a causal role as a restriction on crossing dependencies, but rather emerges as a result of the rarity of crossing dependencies plus low tree depth.

In this work, we evaluate a number of crossing constraints to determine if dependency corpora give evidence for them as true independent constraints. Our **null hypothesis** is that crossing dependencies are formally unrestricted, but occur at a certain low rate per dependency arc. The alternative to the null hypothesis is the **true constraint hypothesis** (TCH), which is that there is a real dispreference for crossing dependencies violating that specific constraint, arising from grammar or cognitive pressures.

We compare the TCH against the null hypothesis by comparing natural language dependency trees with randomly generated baseline trees. The baseline trees simulate the null hypothesis: they consist of randomly generated trees where crossing dependencies have been inserted randomly at the same overall rate per dependency as in the real trees, but with no formal restrictions (more on this in Section 3.2). If the distribution of gap degree, edge degree, etc., in random baseline trees is indistinguishable from real language trees, then we cannot reject the null hypothesis: in that case dependency corpora would not show evidence for the TCH. On the other hand, if a formal measure like gap degree is minimized in observed data over the random baseline, then this is evidence against the null hypothesis and for the TCH.

Our paper is organized as follows. In Section 2, we review the crossing constraints that we will test. In Section 3, we discuss the natural language dataset and the random baselines. We present the results in Section 4. Section 5 concludes.

2 Measures

In order to test the TCH, we compare the distributions of violations of crossing constraints in random baseline trees vs. real language trees. Below we discuss the crossing constraints used in our investigation. In addition, we also discuss the properties of the dependency tree that are used in our comparison of real vs. random trees. In particular, we will be testing whether the correlation between these dependency tree properties and crossing constraint violations is the same in real vs. random trees.

2.1 Crossing Constraints

Gap degree: The **gap degree** of a node X is the number of discontinuities in the projection of node X. For example, in Figure 2, the projection chain of node X_h contains two discontinuities; these discontinuities are present in $X_h \rightarrow X_d$ and in $X_d \rightarrow X_g$. Therefore, the gap degree of node X_h is 2. On the other hand, the gap degree of node X_d is 1. The gap degree of a dependency tree is the maximum among the gap degrees of its nodes (Kuhlmann and Nivre, 2006). In Figure 2, the gap degree of the tree is 2 as the highest gap degree (associated with X_h) is 2. Since gap degree is number of discontinuities in a projection chain, it is upper bounded by the length of projections chains.

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 3. The **edge degree** of a dependency arc $X_h \rightarrow X_d$ is the number of nodes in the span e which are neither dominated by some node in the span e nor dominated by the head X_h. For example, arc $X_h \rightarrow X_d$ in Figure 3(a) and 3(b) has an edge degree of 2 because node X_i and X_b are not dominated by any node in the span e. In addition, they are also not dominated by 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 edge degree might be limited in natural language. From an online processing perspective, higher edge degree in a subtree results in the 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 crossing: The number of **end-point crossings** is the number of heads which dominate the gap of an arc. Given an arc $X_h \rightarrow X_d$ with a span e containing X_i, X_a and X_b as in Figure 3, the end-point crossing of arc $X_h \rightarrow X_d$ is defined as the number of heads modified by the nodes in e that are not part of the projection chain of X_h. For example, in Figure 3(a) and 3(b), X_i and X_b are not part of the projection chain of X_h, in other words they are not dominated by either X_h or any node in the span e. In 3(a), 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 3(b), 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 et al., 2013). Pitler et al. (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 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.

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

(b) : Edge degree=2, End-point crossing=2

Figure 3: 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. 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.

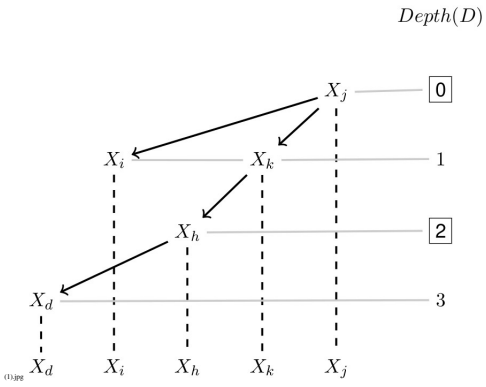

Figure 4: A schematic diagram for heads' depth difference (HDD).

Heads' depth difference (HDD): For a crossing dependency $X_h \rightarrow X_d$, suppose that X_i is the node which creates discontinuity, i.e. X_i is not directly or indirectly dominated by X_h (see Figure 4). 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) is defined as the difference between the depth of head of the crossing dependency X_h and depth of head of the intervener X_j. This is schematically shown in

Figure 4. 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 $\text{depth}(X_h)$ is the hierarchical position of the head of the non-projective dependency (X_h) and $\text{depth}(X_j)$ 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 et al., 1999). However, increasing HDD seems to correlate with increased online processing difficulty for humans (Phillips et al., 2005). More generally, HDD has been proposed (see Yadav et al., 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.

2.2 Dependency tree properties

We study violations of crossing constraint as a function of the following properties of dependency trees.

Sentence length: Sentence length is measured as the total number of nodes in a dependency tree.

Arity: The arity of a node is the total number of dependents of that node. We quantify arity as a global property of a tree by taking the maximum arity per node in the tree.

Tree depth: Tree depth is the number of heads in the longest projection chain in a dependency tree (see Figure 2). Tree depth represents the maximum number of levels of embedding occurring in a tree.

3 Data and methodology

3.1 Natural languages dataset

We use the Universal Dependencies (UD v2.3) treebanks of 14 languages as a dataset (Nivre et al., 2018). The languages were selected for typological diversity: the dataset contains 8 head-initial languages and 6 head-final languages. We do not include dependencies marking punctuation (labeled as 'punct' in UD scheme) and the abstract root of the tree (labeled as 'root' in UD scheme) in our analysis.

As we discuss below, the process of sampling random baseline trees makes it prohibitively difficult to study all languages in the UD dataset. Therefore we study treebanks of 14 languages: German, English, Hindi, French, Arabic, Russian, Czech, Italian, Spanish, Afrikaans, Japanese, Korean, Bulgarian and Slovak. We present results aggregating over dependency trees from all these languages.

3.2 Random baseline

Our null hypothesis is that the only restriction on crossing dependencies is that they are rare, i.e. that they occur at some certain low rate per dependency in a sentence. We instantiate the null hypothesis by sampling random trees which are constrained to have the same distribution over sentence length and number of crossings per dependency as a corpus of some natural language.

We control for sentence length and crossing rate in the random trees in the following way. For each real dependency tree t of length n in a corpus, we sample random trees t' from a uniform distribution over n^{n-1} directed labeled tree structures with n nodes using Prüfer codes (Prüfer, 1918). We control for the crossing rate by rejection sampling: we reject random samples t' which do not have the same number of crossings as the original tree t. For long sentences (over length 12), the rejection sampling process is prohibitively slow, because the vast majority of random trees for $n \geq 12$ have a very large number of crossings. So in the present work we only consider sentences of length less than 12.

Since we are only controlling the number of crossings and the sentence length, the distribution of arity and depth in random baseline trees is quite different from real language trees. In particular, we find that

the growth of tree depth with respect to sentence length is faster for random baseline trees. In addition, the growth of arity with sentence length in the random tree is slower. In sum, random baseline trees are typically deeper than real trees.

3.3 Testing the Null and True Constraint Hypotheses

We compare the rate at which crossing constraints are violated in real trees as compared with random baseline trees, as a function of sentence length, arity, and tree depth. We evaluate the difference between real and random trees statistically using mixed-effects Poisson regression (Gelman and Hill, 2007; Baayen et al., 2008). We fit the regression to predict the rate of constraint violations as a function of dependency tree features (length, depth, and arity) and a dummy-coded variable encoding whether a given tree is real or random. We also include by-language random intercepts. For example, we predict the gap degree g_i of the ith sentence s_i in the jth language as:

$$\log E[g_i] = \beta_0 + \beta_l |s_i| + \beta_r r_i + \beta_{lr} r_i |s_i| + \gamma_j + \varepsilon, \tag{2}$$

where $|s_i|$ is the length of sentence s_i, r_i is an indicator variable with value 1 for a real tree and 0 for a baseline tree, and γ_j, subject to L_2 regularization, is a random intercept for the jth language. The fitted value of the **interaction coefficient** β_{lr} gives the extent to which the growth rate of gap degree as a function of sentence length differs between the real and the random trees. If β_{lr} is significantly negative, then this would mean that gap degree grows more slowly with sentence length in real trees as compared with random trees, i.e. gap degree would be minimized in real trees.

4 Results

We compared the regression pattern of each measure with length, arity and depth between observed and random baseline trees. Below we report the results for each crossing constraint. A summary of all regression results is found in Table 1.

4.1 Gap degree

We find that the distribution of gap degree as a function of sentence length and arity is not significantly different between real and random trees (see Figure 5). In particular, the interaction between length/arity and tree type was not significant in the respective models (see Table 1). However, growth rate of gap degree with tree depth is significantly different between real and random trees ($p < .001$). In other words, we found no evidence for the TCH for gap degree as a function of length and arity: the distribution of gap degree in natural language trees can be fully explained without formal restrictions on crossing dependencies or tree structures. However, the results with respect to depth provide support for the TCH.

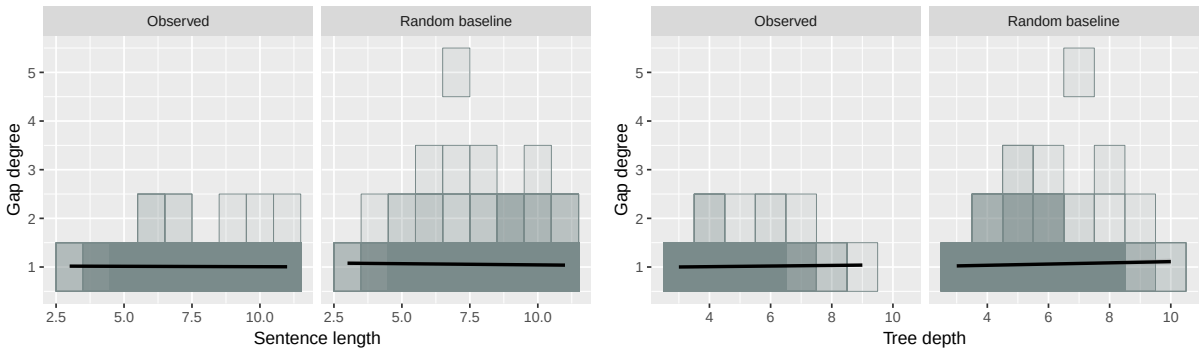

Figure 5: Gap degree as a function of sentnece length and tree depth in real and random trees. In this and all other figures, for visual clarity, we only display results for trees with at least one crossing dependency. All statistical tests are performed using all trees.

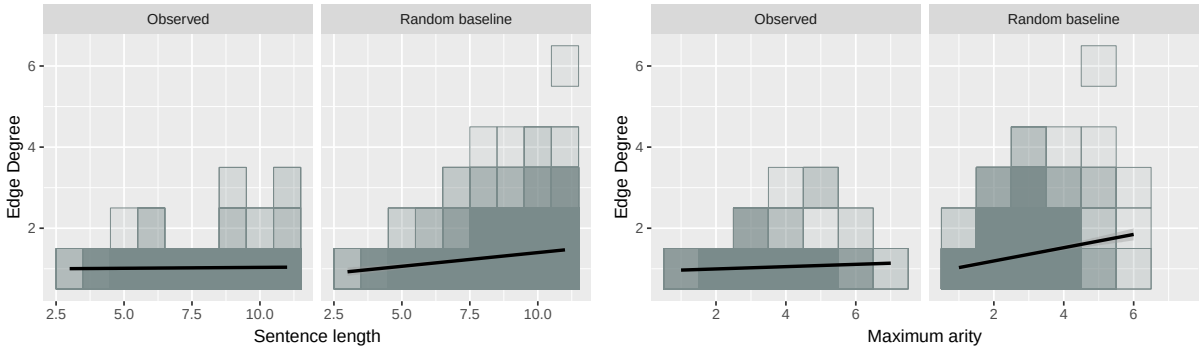

Figure 6: Edge degree as a function of sentence length and tree maximum arity for real and random trees.

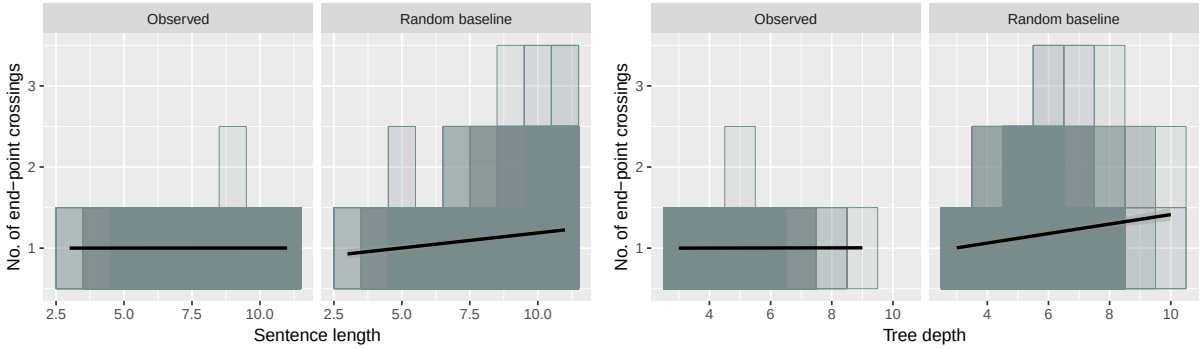

Figure 7: End-point crossings as a function of sentence length and tree depth in real and random trees.

4.2 Edge degree

As shown in Figure 6, edge degree grows faster in random trees in comparison to real trees as a function of sentence length, arity and depth. The mixed-effects Poisson regression models show that the three interaction coefficients (for length, arity, and depth) are significant in the respective models (see Table 1). This provides evidence for the TCH for edge degree.

4.3 End-point crossings

As shown in Figure 7, we find that end-point crossings grow at a slower rate in real trees as a function of tree depth as compared with random baselines. The results support the TCH for end-point crossings. Similar to gap degree, end-point crossing as a function of maximum arity and sentence length does not differ significantly between real and random trees (see Table 1).

4.4 Heads' Depth Difference (HDD)

The results show that HDD decreases with sentence length in real trees, and the rate of decrease is less than in random trees (Figure 8). HDD is also much higher in random trees compared to real trees as a function of tree depth. These results support the TCH for HDD. HDD does not differ between real and random tree with respect to maximum arity (see Table 1).

5 Conclusion

We found that the distribution of gap degree, edge degree, end-point crossing and HDD cannot be explained solely in terms of sentence length and the rate of crossings. These constraints are violated at a different rate as a function of various tree properties than would be expected in random trees, suggesting that they may constitute real formal restrictions on trees.

The results show that the behavior of these crossing constraints differ depending dependency tree properties. Gap-degree and end-point crossings in real vs. random trees are only different as a function

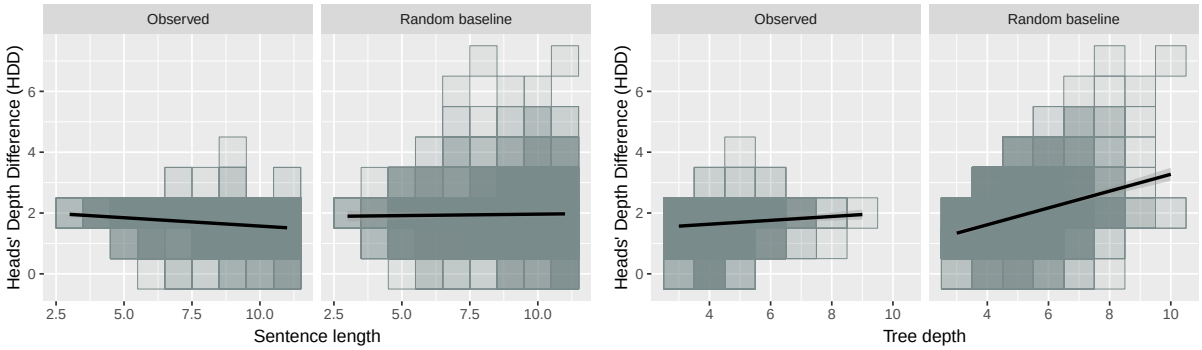

Figure 8: HDD as a function of sentence length and tree depth in real and random trees.

of tree depth (which itself has a very different distribution between real and random trees). HDD in real vs random trees is indistinguishable as a function of arity, but is different for tree depth and sentence length. Edge degree, on the other hand, emerges as the crossing constraint that is most distinct between real and random trees: its distribution is significantly different as a function of all three tree properties.

Our results do not rule out the possibility that the correlations reported here might themselves be epiphenomenal, resulting from other graph-theoretic properties of real dependency trees which were not controlled for here. For example, a great deal of work has shown that syntactic dependency trees are subject to dependency length minimization: a pressure for the linear distance between syntactic heads and dependents to be short (Hawkins, 1994; Gibson, 1998; Liu, 2008; Futrell et al., 2015) (for recent reviews, see Liu et al., 2017; Temperley and Gildea, 2018), and this pressure has been argued to underly the scarcity of crossing dependencies in general (Ferrer-i-Cancho, 2006; Ferrer-i-Cancho and Gómez-Rodríguez, 2016; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). It is also possible that the differences between real trees and random trees in our results are driven by differences in the depth and arity of these trees, or by UD annotation decisions such as the use of content-head dependencies.

Our work provides a strong framework for evaluating any such theory that aims to predict the particular distribution of crossing dependencies in natural language. A syntactic theory can be tested in our framework by creating random baselines that control for the stipulations of the theory and then statistically comparing the distribution of crossing constraint violations with real trees. To that end, we make the code for our analysis freely available at `http://github.com/yadavhimanshu059/measures_of_nonProjectivity`.

Acknowledgments

The authors thank Tim O'Donnell for helpful discussion and the anonymous reviewers for helpful comments on the paper.

References

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

Emmon Bach, Colin Brown, and William D. Marslen-Wilson. 1986. Cross and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 1(4):249–262.

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

Noam Chomsky. 2007. Approaching UG from below. In *Interfaces + recursion = language?: Chomsky's Minimalism and the view from syntax-semantics*, pages 1–29. Mouton de Gruyter.

Ramon Ferrer-i-Cancho. 2006. Why do syntactic links not cross? *Europhysics Letters*, 76(6):1228.

Dependent variable	Independent variable	β Estimate	Std. Error	p value	
Gap degree	Sentence length	0.75	0.04	$<$ 2e-16	*
	Observed	-0.07	0.05	0.174	n.s.
	Sentence length $\times$ Observed	-0.03	0.05	0.563	n.s.
	Arity	0.25	0.03	2.88e-14	*
	Observed	-0.18	0.04	0.00013	*
	Arity $\times$ Observed	-0.06	0.04	0.1570	n.s.
	Depth	0.52	0.02	$<$ 2e-16	*
	Observed	0.24	0.05	1.23e-05	*
	Depth $\times$ Observed	0.29	0.04	2.43e-10	*
Edge degree	Sentence length	0.37	0.03	$<$ 2e-16	*
	Observed	-0.20	0.04	1.41e-06	*
	Sentence length $\times$ Observed	-0.11	0.04	0.0153	*
	Arity	0.09	0.02	0.0015	*
	Observed	-0.24	0.04	3.65e-09	*
	Arity $\times$ Observed	-0.13	0.04	0.0009	*
	Depth	0.32	0.02	$<$ 2e-16	*
	Observed	0.04	0.04	0.33	n.s.
	Depth $\times$ Observed	0.21	0.04	1.02e-06	*
End-point crossing	Sentence length	0.32	0.03	$<$ 2e-16	*
	Observed	-0.10	0.04	0.0173	*
	Sentence length $\times$ Observed	-0.07	0.04	0.1013	n.s.
	Arity	-0.001	0.03	0.9900	n.s.
	Observed	-0.10	0.04	0.0141	*
	Arity $\times$ Observed	-0.07	0.04	0.1098	n.s.
	Depth	0.34	0.02	$<$ 2e-16	*
	Observed	0.16	0.04	0.0002	*
	Depth $\times$ Observed	0.20	0.04	3.92e-06	*
HDD	Sentence length	0.27	0.02	$<$ 2e-16	*
	Observed	-0.14	0.03	8.78e-06	*
	Sentence length $\times$ Observed	-0.08	0.03	0.0152	*
	Arity	-0.11	0.02	7.36e-06	*
	Observed	-0.10	0.03	0.0025	*
	Arity $\times$ Observed	-0.02	0.03	0.4695	n.s.
	Depth	0.44	0.02	$<$ 2e-16	*
	Observed	0.21	0.03	1.19e-08	*
	Depth $\times$ Observed	0.13	0.03	8.78e-06	*

Table 1: Mixed-effect Poisson regression results for all the crossing constraints and dependency tree measures for 14 languages. "Observed" is an indicator variable with value 1 for observed trees and 0 for random trees, the same as r_i in Equation 2. A significant interaction between an independent variable and Observed rejects the null hypothesis.

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

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. https://doi.org/10.1073/pnas.1502134112 Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–

76.

Edward Gibson 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.

Carlos Gómez-Rodríguez and Ramon Ferrer-i-Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96.

John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge.

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

Aravind K. Joshi. 1985. Processing of sentences with intrasentential code switching. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 190–205. Cambridge University Press, Cambridge.

Marco Kuhlmann. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 507–514, Sydney, Australia. Association for Computational Linguistics.

Roger P. Levy, Evelina Fedorenko, Mara Breen, and Edward Gibson. 2012. The processing of extraposed structures in English. *Cognition*, 122(1):12–36.

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

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

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*.

Solomon Marcus. 1965. Sur la notion de projectivité. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 11(2):181–192.

Jens Michaelis. 1998. Derivational Minimalism is mildly context-sensitive. In *Logical Aspects of Computational Linguistics*, volume 98, pages 179–198. Springer.

Joakim Nivre. 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*, Rochester, New York. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, et al. 2018. http://hdl.handle.net/11234/1-2895 Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

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

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

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

Hiroyuki Seki, Takashi Matsumara, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229.

Stuart M. Shieber. 1985. Evidence against the context-freeness of natural language. In *The Formal complexity of natural language*, pages 320–334. Springer.

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič, Jr. 2015. Parsing universal dependency treebanks using neural networks and search-based oracle. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 208–220, Warszawa, Poland. IPIPAN.

David Temperley and Dan Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.

Carl Vogel, Ulrike Hahn, and Holly Branigan. 1996. Cross-serial dependencies are not hard to process. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 157–162. Association for Computational Linguistics.

Himanshu Yadav, 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)*, Pisa,Italy. Linköping University Electronic Press.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. under review. Word order typology interacts with linguistic complexity: a cross-linguistic corpus study.

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

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