# Crossings as a side effect of dependency lengths

Ramon Ferrer-i-Cancho[1,*] & Carlos Gómez-Rodríguez[2]

[1] Complexity & Qualitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega. Jordi Girona Salgado 1-3. 08034 Barcelona, Catalonia, Spain
[2] LyS Research Group, Departamento de Computación, Facultade de Informática, Universidade da Coruña, Campus de A Coruña, 15071 A Coruña, Spain

∗ To whom correspondence should be addressed. E-mail: rferrericancho@cs.upc.edu.

**Abstract**

The syntactic structure of sentences exhibits a striking regularity: dependencies tend to not cross when drawn above the sentence. Here we investigate two competing hypotheses for the origins of non-crossing dependencies. The traditional hypothesis is that the low frequency of dependency crossings arises from an independent principle of syntax that reduces crossings practically to zero. An alternative to this view is the hypothesis that crossings are a side effect of dependency lengths. According to this view, sentences with shorter dependency lengths should tend to have fewer crossings. We recast the traditional view as a null hypothesis where one of the variables, i.e. the number of crossings, is mean independent of the other, i.e. the sum of dependency lengths. The alternative view is then a positive correlation between these two variables. In spite of the rough estimation of dependency crossings that this sum provides, we are able to reject the traditional view in the majority of languages considered. The alternative hypothesis can lead to a more parsimonious theory of syntax.

## 1 Introduction

The syntactic structure of sentences exhibits a striking regularity that was reported in the 1960s: dependencies tend to not cross when drawn above the sentence (Lecerf, 1960; Hays, 1964), as shown in Fig. 1. Here we investigate two competing hypotheses for the origins of non-crossing dependencies.

The traditional hypothesis is that the low frequency of dependency crossings arises from an independent principle of syntax that reduces crossings practically to zero. This view is held by theories of grammar were crossings are not allowed (Sleator & Temperley, 1993; Hudson, 2007; Tanaka, 1997; Kurohashi & Nagao,

1997) and also by parsing frameworks where non-crossing dependencies are not allowed or subject to hard constraints (Sleator & Temperley, 1993; Nivre, 2003; Carreras, 2007; Zhang & Nivre, 2011; Chen & Manning, 2014; Dyer, Ballesteros, Ling, Matthews, & Smith, 2015). It is also shared by research on dependency length minimization where annotations with crossings are discarded (Gildea & Temperley, 2010) and actual dependency lengths are compared with two kinds of baselines where crossings are not allowed or are subject to hard constraints: random orderings and optimal dependency lengths (Temperley, 2008; Liu, 2008; Park & Levy, 2009; Gildea & Temperley, 2010; Futrell, Mahowald, & Gibson, 2015; Gulordava & Merlo, 2015). The traditional view is convenient for simplicity and computational reasons: efficient algorithms for non-crossing dependencies or limited violations are available (Hochberg & Stallmann, 2003; Gildea & Temperley, 2007; Park & Levy, 2009) and is justified by the low frequency of crossings in real languages (Temperley, 2008; Gildea & Temperley, 2010).

An alternative to this view is the hypothesis that crossings are a side effect of dependency lengths (Ferrer-i-Cancho, 2014, 2015). This hypothesis predicts that dependencies should tend to not cross, combining a tendency for shorter dependency lengths to have less crossings and the fact that dependencies are actually short. This challenges the dogma that unconstrained dependency length minimization "does not take into account constraints of projectivity or mild context-sensitivity" (Park & Levy, 2009); and is coherent with the trends towards diachronic reduction of the proportion of crossings in conjunction with dependency length minimization that have recently been observed on Latin and Ancient Greek (Gulordava & Merlo, 2015).

Here we will evaluate these two hypotheses making emphasis on the validity of the traditional view. We will formalize the traditional view as a null hypothesis and the alternative view as an alternative hypothesis. With the help of a collection of dependency treebanks of thirty different languages, we will show that the null hypothesis of the traditional view is rejected for a large majority of treebanks.

## 2 Formalization of the problem

The traditional view can be recast as a simple model that predicts, given a sentence, a zero number of crossings. This deterministic model with no parameter can be generalized as a stochastic model with one parameter $a$ that defines the expected number of crossings. Suppose that $C$ is the number of crossings of a sentence and $E[C|sentence]$ is the expectation of $C$ over all possible orderings of a sentence (Ferrer-i-Cancho, 2015). Then the traditional view can be defined as

$$E[C|sentence] = a. \tag{1}$$

where $a$ is a constant such that $a \geq 0$. $a = 0$ implies a ban of crossings because $C \geq 0$. The parameter $a$ allows one to model crossings in languages with varying frequencies of crossings (from languages where there are no crossings to languages where crossings occur with a certain frequency). If the relevant

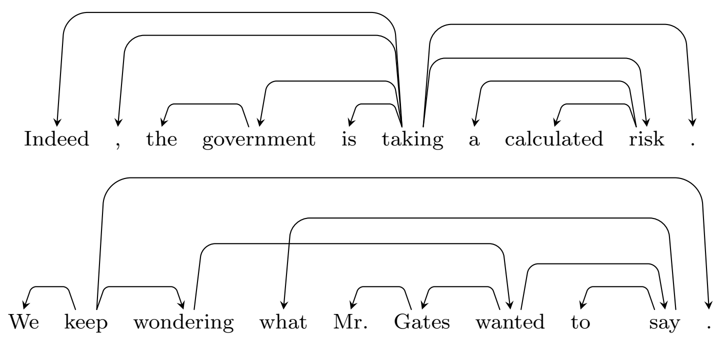


Figure 1: A sentence with no crossings (top) and a sentence with one crossing (bottom), both taken from the subset of the English Penn Treebank with Stanford dependency annotations included in HamleDT 2.0 (Rosa et al., 2014)

information of a sentence is $D$, the sum of dependency lengths, the alternative hypothesis can be modeled simply as (Ferrer-i-Cancho, 2015)

$$E[C|D] = g(D) \tag{2}$$

and then the traditional hypothesis can be written as

$$E[C|D] = a. \tag{3}$$

A limitation of $E[C|D]$ is that it is defined over a set of possible linearizations that includes some that are very unlikely, cognitively harder or "ungrammatical". In this article, we focus on real linearizations and therefore we consider $E_{TB}[C|D]$, the expectation of $C$ given $D$ over the ensemble of linearizations of the sentences of a treebank (TB). $E_{TB}[C|D]$ needs to be refined: the distribution of $D$ depends on the length of the sentence and then values of $D$ from sentences of different length should not be mixed (Ferrer-i-Cancho & Liu, 2014). The same kind of problem is also likely to concern $C$. For this reason, instead of $E_{TB}[C|D]$, we choose $E_{TB}[C|n, D]$, i.e. the expectation of $C$ conditioning on sentences of the treebank that have length $n$ and their sum of dependency lengths is $D$. The traditional view is defined simply as

$$E_{TB}[C|n, D] = a_{TB}(n), \tag{4}$$

where $a_{TB}$ is a constant with respect to $D$ that depends on $n$, such that $a_{TB}(n) = 0$ for $n < 4$ (since $C = 0$ in this case (Ferrer-i-Cancho, 2013a)) and $a_{TB}(n) = E_{TB}[C|n]$ (Ferrer-i-Cancho, Hernández-Fernández, Baixeries, Dębowski, & Mačutek, 2014). The latter allows one to formulate the traditional view equivalently as

$$E_{TB}[C|n, D] = E_{TB}[C|n]. \tag{5}$$

Thus, given a treebank and a sentence length $n$, the traditional hypothesis predicts that a sentence will have, on average, a number of crossings that coincides with the mean number of crossings of the sentences of length $n$. Accordingly, the alternative view is modeled by

$$E_{TB}[C|n, D] = g_{TB}(n, D), \tag{6}$$

where $g_{TB}(n, D)$ is a strictly monotonically increasing function of $D$ when $n$ remains constant. In this article, we want to remain agnostic about the exact mathematical form of $g_{TB}(n, D)$. Our focus is on the validity of the traditional view. Concerning the alternative view, we are only interested in the sign of the correlation between $C$ and $D$. A positive correlation provides support for the hypothesis that crossings are side effect of dependency lengths.

Eq. 5 is interesting because it indicates that the traditional view is equivalent to $C$ being mean independent of $D$ when $n$ is given, in the language of probability theory (Poirier, 1995, p. 67). From the perspective of statistical hypothesis testing, the traditional view is a null hypothesis (mean independence), while the alternative view (a positive correlation between $D$ and $C$) is an alternative hypothesis.

Fig. 2 compares the relationship between $D$ and $C$ in sentences of length 18 in an English dependency treebank. In this case, the traditional view is

$$E_{TB}[C|n, D] = a_{TB}(n) \tag{7}$$

with $a_{TB}(n) = 0.08$, the mean number of crossings in sentences of length 18 in that treebank. This very low number casts doubts on the adequacy of the null hypothesis for the large values of $C$ that are found especially for large values of $D$ in Fig. 2. The Kendall $\tau$ correlation between $C$ and $D$ is $\tau = 0.03$ (p-value = 0.28) indicating a weak but positive tendency of $C$ to increase as $D$ increases. In this article, we will study collections of sentences with syntactic dependency annotations (treebanks) of different languages, to check if the number of positive $\tau$ correlations across sentence lengths is significantly high. If that happens, we will conclude that an autonomous bound on crossings (Eq. 4) does not hold in general for that treebank. We would like to emphasize that the goal of this article is not to predict the actual number of crossings with great accuracy as in related work (Ferrer-i-Cancho, 2014) but examining the validity of the traditional view with a simple (and statistically sound) approach. $D$ is a rough predictor of crossings because the probability that two dependencies cross is determined by their individual lengths and whether they share vertices or not (Ferrer-i-Cancho, 2014, 2015). $D$ can be seen as a lossy compression of the dependency lengths of a sentence into a single value. Thus, it is rather surprising that the rough predictions that $D$ offers allow us to reject the traditional view in the majority of treebanks, as we will see.

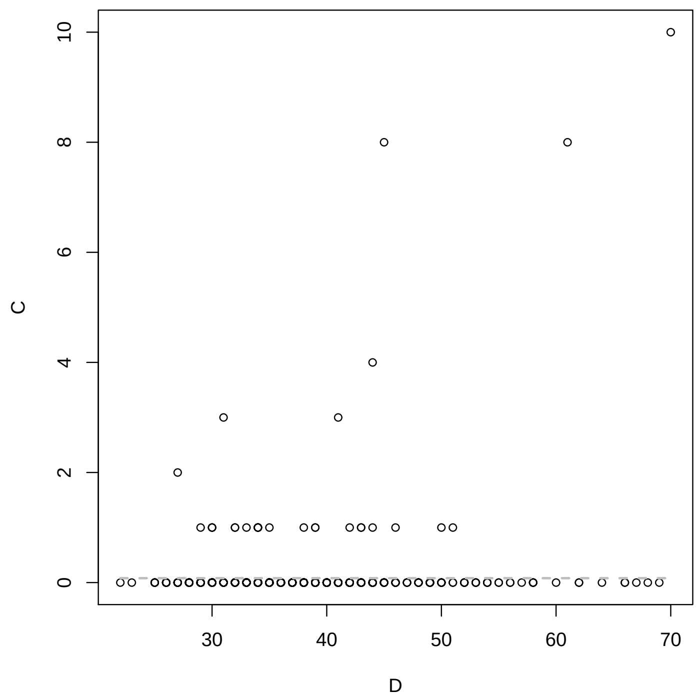


Figure 2: Crossings ($C$) versus sum of dependency lengths ($D$) in sentences of length 18 in an English treebank (we use Prague dependencies from HamleDT 2.0, see Section 3.1). 18 is the typical sentence length in this treebank. The average prediction made by the null hypothesis is also shown (gray dashed line).

# 3 Materials and methods

## 3.1 Materials

We employ HamleDT 2.0, a collection of dependency treebanks of 30 different languages (Rosa et al., 2014). The collection provides sentences with syntactic dependency annotations following two different criteria: Prague dependencies (Hajič et al., 2006) and Stanford dependencies (Marneffe et al., 2014). This collection allows one to explore a set of typologically diverse languages and control for the effect of annotation criteria.

Each syntactic dependency structure in the treebanks was preprocessed by removing nodes corresponding to punctuation tokens. To preserve the syntactic structure of the rest of the nodes, non-punctuation nodes that had a punctuation node as their head were attached as dependents of their nearest non-punctuation ancestor. Null elements, which appear in the Bengali, Hindi and Telugu corpora, were also subject to the same treatment as punctuation.

After this preprocessing, syntactic dependency structures that did not define a tree were removed. The reason is that we wanted to avoid the statistical problem of mixing trees with other kinds of graphs, e.g. the potential number of crossings depends on the number of edges (Ferrer-i-Cancho, 2013a, 2013b, 2015).

## 3.2 Methods

For each every sentence length of a treebank, we want to investigate if the null hypothesis that $C$ is mean independent of $D$ actually holds. This can be tested with the help of the Kendall $\tau$ correlation between $C$ and $D$. Suppose that $c_1$ and $c_2$ are two observations of $C$ and $d_1$ and $d_2$ are two observations of $D$. Then $(c_1, d_1)$ and $(c_2, d_2)$ are said to be concordant if $(c_1 - c_2)(d_1 - d_2) > 0$ (the ranks of both elements agree), and discordant if $(c_1 - c_2)(d_1 - d_2) < 0$ (they disagree). Then, Kendall $\tau$ correlation is defined as (Conover, 1999)

$$\tau = \frac{N_c - N_d}{N_0}, \tag{8}$$

where $N_c$ is the number of concordant pairs, $N_d$ is the number of discordant pairs and $N_0$ is the total number of pairs.

Sentences where $n < 4$ were excluded from the analysis because $C = 0$ for them (Ferrer-i-Cancho, 2013a). Sentence lengths with less than two sentences were removed because $N_0 = 0$ and then $\tau$ is not properly defined. For each treebank, we calculated the Kendall $\tau$ correlation between $D$ and $C$ for every sentence length (after applying all the filtering criteria described in this section). Then we calculated $p(\tau \geq 0)$, the proportion of sentence lengths where $\tau \geq 0$. If $p(\tau \geq 0)$ is sufficiently high then the null hypothesis of mean independence is rejected. The significance of $p(\tau \geq 0)$ was determined with the help of a Monte Carlo method that takes as input the vectors $\vec{D^n} = \{d_1^n, \ldots, d_i^n, \ldots, d_m^n\}$

and $\vec{C^n} = \{c_1^n, \ldots, c_i^n, \ldots, c_m^n\}$ of every sentence length $n$ ($d_i^n$ and $c_i^n$ are, respectively, the sum of dependency lengths and number of crossings of the $i$-th sentence of length $n$). This method consists of generating $T$ randomizations of the input vectors and estimating the p-value of the test as the proportion of times that $p_c(\tau \geq 0) \geq p(\tau \geq 0)$, where $p_c(\tau \geq 0)$ is the value of $p(\tau \geq 0)$, over $T$ randomizations of the vectors. A randomization consists of replacing the vector $\vec{D^n}$ for each sentence length with a uniformly random permutation. For this article, we use $T = 10^4$ and a significance level of 0.05.

We could have determined the significance of $p(\tau \geq 0)$ by means of a binomial test: under the assumption of independence between $D$ and $C$ and assuming that there are no ties among values, the probability that $\tau \geq 0$ is $1/2$ (Prokhorov, 2001). However, ties of $C$ abound (many sentences have $C = 0$, see also Table 2). For this reason, the Monte Carlo test above yields a more accurate estimation of the true p-value.

It is convenient to split $p(\tau \geq 0)$ as $p(\tau > 0) + p(\tau = 0)$ and inspect $p(\tau = 0)$ because Kendall $\tau = 0$ is due to $N_c = N_d$ (recall Eq. 8). High p-values of $p(\tau \geq 0)$ could be due to high $p(\tau = 0)$, which in turn would be due to $C = 0$ for many sentence lengths. To see it, consider the following extreme case: a treebank where $C = 0$ in all sentences. In that case, $N_c = N_d = 0$ for all sentence lengths and then $p(\tau \geq 0) = p(\tau = 0)$. Interestingly, $\tau$ would remain zero for all sentence lengths after randomization and then the p-value of the Monte Carlo test would be 1. That has been the case of the Romanian treebank with Prague dependencies (Tables 1 and 2).

# 4 Results

Table 1 shows that $p(\tau \geq 0)$ is significantly high in about three fourths of the languages for Prague dependencies (eight treebanks have a p-value above the significance level) and to a much larger extent for the Stanford dependencies (only five treebanks have a p-value above the significance level). Thus, there is a minority of languages where there is not enough support for the hypothesis that crossing dependencies are a side effect of dependency lengths. Interestingly, $p(\tau = 0)$ is especially high in the treebanks where $p(\tau \geq 0)$ is not significantly high. A possible explanation for the failure of the alternative view in those treebanks is that $C = 0$ in the majority of sentence lengths. Let us call $p_0$ the proportion of sentence lengths where all sentences have $C = 0$. Table 2 indicates that the five treebanks where $p(\tau \geq 0)$ is not significantly high for Stanford dependencies coincide with the five treebanks with the largest $p_0$. The situation for Prague dependencies is similar: the top six largest values of $p_0$ are taken by six treebanks where $p(\tau \geq 0)$ is not significantly high.

| Treebank | Prague | | | | Stanford | | | |
|---|---|---|---|---|---|---|---|---|
| | $M$ | $p(\tau = 0)$ | $p(\tau > 0)$ | p-value | $M$ | $p(\tau = 0)$ | $p(\tau > 0)$ | p-value |
| Arabic | 90 | 0.38 | 0.34 | 0.239 | 90 | 0.067 | 0.7 | 0.0001 |
| Basque | 33 | 0.091 | 0.85 | $< 10^{-4}$ | 33 | 0.03 | 0.82 | 0.0001 |
| Bengali | 16 | 0.19 | 0.38 | 0.7372 | 17 | 0.29 | 0.53 | 0.0871 |
| Bulgarian | 51 | 0.078 | 0.69 | 0.0006 | 52 | 0.077 | 0.77 | $< 10^{-4}$ |
| Catalan | 86 | 0.16 | 0.76 | $< 10^{-4}$ | 86 | 0.047 | 0.72 | $< 10^{-4}$ |
| Czech | 73 | 0.027 | 0.78 | $< 10^{-4}$ | 74 | 0.054 | 0.76 | $< 10^{-4}$ |
| Danish | 56 | 0.11 | 0.62 | 0.005 | 57 | 0.07 | 0.77 | $< 10^{-4}$ |
| Dutch | 52 | 0.019 | 0.81 | $< 10^{-4}$ | 52 | 0 | 0.85 | $< 10^{-4}$ |
| English | 66 | 0.11 | 0.64 | 0.0001 | 66 | 0.045 | 0.64 | 0.0089 |
| Estonian | 22 | 0.82 | 0.14 | 0.3027 | 22 | 0.45 | 0.14 | 0.9877 |
| Finnish | 33 | 0.15 | 0.79 | $< 10^{-4}$ | 33 | 0.091 | 0.88 | $< 10^{-4}$ |
| German | 72 | 0.042 | 0.71 | $< 10^{-4}$ | 72 | 0.014 | 0.64 | 0.0151 |
| Greek(ancient) | 53 | 0 | 0.94 | $< 10^{-4}$ | 53 | 0 | 0.89 | $< 10^{-4}$ |
| Greek(modern) | 63 | 0.24 | 0.49 | 0.0358 | 64 | 0.14 | 0.66 | 0.0001 |
| Hindi | 58 | 0.069 | 0.78 | $< 10^{-4}$ | 58 | 0.086 | 0.74 | $< 10^{-4}$ |
| Hungarian | 62 | 0.032 | 0.74 | $< 10^{-4}$ | 61 | 0.049 | 0.64 | 0.0048 |
| Italian | 59 | 0.34 | 0.53 | 0.0002 | 59 | 0.12 | 0.73 | $< 10^{-4}$ |
| Japanese | 37 | 0.97 | 0 | 1 | 37 | 0 | 0.95 | $< 10^{-4}$ |
| Latin | 46 | 0 | 0.72 | 0.0042 | 46 | 0.043 | 0.8 | $< 10^{-4}$ |
| Persian | 71 | 0.028 | 0.25 | 0.9999 | 72 | 0.042 | 0.76 | $< 10^{-4}$ |
| Portuguese | 79 | 0.063 | 0.71 | $< 10^{-4}$ | 79 | 0.089 | 0.75 | $< 10^{-4}$ |
| Romanian | 38 | 1 | 0 | 1 | 38 | 0.21 | 0.5 | 0.1266 |
| Russian | 66 | 0.076 | 0.76 | $< 10^{-4}$ | 65 | 0.031 | 0.83 | $< 10^{-4}$ |
| Slovak | 74 | 0.068 | 0.64 | 0.0008 | 76 | 0.026 | 0.75 | $< 10^{-4}$ |
| Slovene | 46 | 0.087 | 0.59 | 0.027 | 49 | 0.041 | 0.73 | 0.0001 |
| Spanish | 81 | 0.16 | 0.79 | $< 10^{-4}$ | 80 | 0.037 | 0.84 | $< 10^{-4}$ |
| Swedish | 59 | 0.1 | 0.81 | $< 10^{-4}$ | 61 | 0.033 | 0.75 | $< 10^{-4}$ |
| Tamil | 31 | 0.84 | 0.13 | 0.2015 | 31 | 0.68 | 0.26 | 0.0514 |
| Telugu | 8 | 0.88 | 0.12 | 0.5315 | 8 | 0.62 | 0.38 | 0.1727 |
| Turkish | 43 | 0.07 | 0.67 | 0.0033 | 44 | 0.11 | 0.75 | $< 10^{-4}$ |

Table 1: Summary of the analysis of the correlation between $D$ and $C$. For every treebank, we show the number of different sentence lengths considered ($M$), the proportion of sentence lengths where Kendall $\tau$ is equal or greater than zero ($p(\tau = 0)$ and $p(\tau > 0)$, respectively), and the p-value of the Monte Carlo test for the significance of $p(\tau \geq 0) = p(\tau > 0) + p(\tau = 0)$.

| Prague | | Stanford | |
|---|---|---|---|
| Treebank | $p_0$ | Treebank | $p_0$ |
| **Romanian** | 1 | **Tamil** | 0.68 |
| **Japanese** | 0.97 | **Telugu** | 0.62 |
| **Telugu** | 0.88 | **Estonian** | 0.45 |
| **Tamil** | 0.84 | **Bengali** | 0.29 |
| **Estonian** | 0.82 | **Romanian** | 0.21 |
| **Arabic** | 0.34 | Turkish | 0.11 |
| Italian | 0.32 | Greek(modern) | 0.11 |
| Greek(modern) | 0.22 | Finnish | 0.091 |
| **Bengali** | 0.19 | Hindi | 0.086 |
| Catalan | 0.16 | Italian | 0.051 |
| Spanish | 0.14 | Bulgarian | 0.038 |
| Finnish | 0.12 | Catalan | 0.035 |
| Danish | 0.11 | Arabic | 0.033 |
| Swedish | 0.085 | Basque | 0.03 |
| Turkish | 0.07 | Spanish | 0.025 |
| Russian | 0.061 | Latin | 0.022 |
| Bulgarian | 0.059 | Danish | 0.018 |
| Hindi | 0.052 | Hungarian | 0.016 |
| English | 0.045 | Russian | 0.015 |
| Hungarian | 0.032 | English | 0.015 |
| Basque | 0.03 | Portuguese | 0.013 |
| Portuguese | 0.025 | Czech | 0 |
| Slovene | 0.022 | Dutch | 0 |
| **Persian** | 0.014 | German | 0 |
| Czech | 0 | Greek(ancient) | 0 |
| Dutch | 0 | Japanese | 0 |
| German | 0 | Persian | 0 |
| Greek(ancient) | 0 | Slovak | 0 |
| Latin | 0 | Slovene | 0 |
| Slovak | 0 | Swedish | 0 |

Table 2: $p_0$, the proportion of sentence lengths where $C = 0$ for all sentences. Treebanks are sorted decreasingly by $p_0$. The treebanks where the null hypothesis could not be rejected according to Table 1 appear in boldface.

# 5 Discussion

We have rejected the traditional hypothesis of crossings as being constrained independently from the dependency lengths in a large majority of treebanks (47 out of 60) thanks to a positive correlation between crossings ($C$) and dependency lengths ($D$) that holds across sentence lengths. The fact that the number of rejections depends on the annotation style (eight treebanks for Prague dependencies, five treebanks for Stanford dependencies) suggests that annotation criteria are crucial. Indeed, we have seen that there is a strong tendency for $C = 0$ across sentence lengths in those treebanks (Table 2).

Before concluding prematurely that the minority of languages where the traditional view could not be rejected constitute evidence of an autonomous ban of crossings, we should recall the limited capacity of $D$ to predict crossings (Section 2) and reflect on the influence that syntactic dependency annotation criteria have had on the results due to:

- A belief in a ban of crossings (Sassano, 2005; Iwatate, Asahara, & Matsumoto, 2008) or a principle of dependency length minimization.
- Automatic conversions from phrase structure grammar to dependency treebanks (Bick, Uibo, & Müürisep, 2004; Kawata & Bartels, 2000), where crossings could be less likely with respect to direct annotations based on dependency grammar.
- Annotation by automatic parsing followed by manual revision (Ramasamy & Žabokrtský, 2012), which can be biased due to either the parser not supporting crossings, or just having low recall for crossing dependencies, a common limitation even in modern non-projective parsers (Björkelund & Kuhn, 2012; Gómez-Rodríguez & Nivre, 2013).
- The need of avoiding crossings to facilitate parsing by computers, as treebanks and annotation guidelines are often developed with this goal in mind (Călăcean, 2008; Begum et al., 2008).
- Aesthetical considerations: dependency structures without crossings being considered nicer than structures with crossings.

Given all the preceding considerations, our results and previous work (Ferrer-i-Cancho, 2014, 2015) provide support for the hypothesis that dependency crossings are a side effect of dependency lengths. By not requiring a belief in an autonomous ban of crossings (Temperley, 2008; Park & Levy, 2009; Gildea & Temperley, 2010; Futrell et al., 2015), this hypothesis promises to help develop a more parsimonious theory of syntax.

# 6 Acknowledgments

RFC is funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and also the APCOM project (TIN2014-57226-P) from

MINECO (Ministerio de Economia y Competitividad). CGR is partially funded by the TELEPARES-UDC project (FFI2014-51978-C2-2-R) from MINECO and the grant R2014/034 from Xunta de Galicia.

# References

Begum, R., Husain, S., Dhwaj, A., Misra, D., Bai, L., & Sangal, R. (2008). Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume I* (pp. 721–726). Hyderabad, India: AFNLP.

Bick, E., Uibo, H., & Müürisep, K. (2004). Arborest – a VISL-style treebank derived from an Estonian constraint grammar corpus. In *Proceedings of Treebanks and Linguistic Theories* (pp. 9–20). Tübingen, Germany: University of Tübingen.

Björkelund, A., & Kuhn, J. (2012). Comparing non-projective strategies for labeled graph-based dependency parsing. In *Proceedings of COLING 2012: Posters* (pp. 135–144). Mumbai, India: The COLING 2012 Organizing Committee.

Carreras, X. (2007). Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL 2007* (pp. 957–961). Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740–750). Stroudsburg, PA, USA: Association for Computational Linguistics.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley. (3rd edition)

Călăcean, M. (2008). *Data-driven dependency parsing for Romanian*. Unpublished master's thesis, Uppsala University.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 1: Long papers)* (pp. 334–343). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ferrer-i-Cancho, R. (2013a). Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics*, *25*, 1-21.

Ferrer-i-Cancho, R. (2013b). Random crossings in dependency trees. *http://arxiv.org/abs/1305.4561*.

Ferrer-i-Cancho, R. (2014). A stronger null hypothesis for crossing dependencies. *Europhysics Letters*, *108*, 58003.

Ferrer-i-Cancho, R. (2015). Non-crossing dependencies: least effort, not grammar. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Towards a theoretical framework for analyzing complex linguistic networks* (p. 203-234). Berlin: Springer.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Baixeries, J., Dębowski, Ł., & Mačutek, J. (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, *13*, 633-644.

Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, *5*, 143-155.

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

Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 184–191). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286-310.

Gómez-Rodríguez, C., & Nivre, J. (2013). Divisible transition systems and multiplanar dependency parsing. *Computational Linguistics*, *39*(4), 799–845.

Gulordava, K., & Merlo, P. (2015). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 121–130). Uppsala, Sweden: Uppsala University.

Hajič, J., Panevová, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., et al. (2006). *Prague Dependency Treebank 2.0.* CDROM CAT: LDC2006T01, ISBN 1-58563-370-4. Linguistic Data Consortium.

Hays, D. (1964). Dependency theory: a formalism and some observations. *Language*, *40*, 511-525.

Hochberg, R. A., & Stallmann, M. F. (2003). Optimal one-page tree embeddings in linear time. *Information Processing Letters*, *87*, 59-66.

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

Iwatate, M., Asahara, M., & Matsumoto, Y. (2008). Japanese dependency parsing using a tournament model. In *Proceedings of the 22nd International Conference on Computational Linguistics - volume 1* (pp. 361–368). Stroudsburg, PA, USA: ACL.

Kawata, Y., & Bartels, J. (2000, September 29). Stylebook for the Japanese treebank in Verbmobil. In *Report 240.* Tübingen, Germany: University of Tübingen.

Kurohashi, S., & Nagao, M. (1997). Kyoto university text corpus project (in Japanese). In *Proceedings of the 3rd Annual Meeting of the Association for Natural Language Processing* (p. 115-118). Kyoto, Japan: Association for Natural Language Processing.

Lecerf, Y. (1960). Programme des conflits - modèle des conflits. *Rapport CETIS No. 4*, 1-24. (Euratom)

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

Marneffe, M.-C. de, Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., et al. (2014). Universal Stanford dependencies: a cross-linguistic typology. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4585–4592). Paris, France: European Language Resources Association (ELRA).

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)* (pp. 149–160). Stroudsburg, PA, USA: ACL/SIGPARSE.

Park, Y. A., & Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the 10th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 335–343). Stroudsburg, PA, USA: Association for Computational Linguistics.

Poirier, D. J. (1995). *Intermediate statistics and econometrics: A comparative approach.* Cambridge: MIT Press.

Prokhorov, A. (2001). Kendall coefficient of rank correlation. In M. Hazewinkel (Ed.), *Encyclopedia of mathematics.* Dordrecht, the Netherlands: Kluwer.

Ramasamy, L., & Žabokrtský, Z. (2012). Prague dependency style treebank for Tamil. In *Proceedings of LREC 2012* (pp. 23–25). Paris, France: European Language Resources Association (ELRA).

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., & Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, & J. Mariani (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2334–2341). Paris, France: European Language Resources Association (ELRA).

Sassano, M. (2005). Using a partially annotated corpus to build a dependency parser for Japanese. In *Proceedings of the Second International Joint Conference on Natural Language Processing: Full papers* (pp. 82–92). Berlin-Heidelberg: Springer.

Sleator, D., & Temperley, D. (1993). Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies (IWPT)* (pp. 277–292). Stroudsburg, PA, USA: ACL/SIGPARSE.

Tanaka, H. (1997). Invisible movement in sika-nai and the linear crossing constraint. *Journal of East Asian Linguistics*, *6*, 143–188.

Temperley, D. (2008). Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, *15*(3), 256-282.

Zhang, Y., & Nivre, J. (2011). Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers - volume 2* (pp. 188–193). Stroudsburg, PA, USA: Association for Computational Linguistics.