Length of non-projective sentences: A pilot study using a Czech UD treebank

Ján Mačutek
Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics
Department of Applied Mathematics and Statistics
Slovakia
jmacutek@yahoo.com

Radek Čech
University of Ostrava
Faculty of Arts
Department of Czech Language
Czech Republic
cechradek@gmail.com

Jiří Milička
Charles University in Prague
Faculty of Arts
Institute of Comparative Linguistics, and
Institute of the Czech National Corpus
Czech Republic
jiri@milicka.cz

Abstract

Lengths (in words) of projective and non-projective sentences from a Czech UD dependency treebank are compared. It is shown that non-projective sentences are significantly longer (in addition, the same result was obtained in this study also for Arabic, Polish, Russian, and Slovak). The hyperpascal distribution, which was suggested as the model for frequency distribution of sentence length measured in words, fits well the data from both projective and non-projective sentences; however, its parameters attain different values for the two groups. Proportions of non-projective sentences in the treebanks used are presented, together with a discussion on factors which can influence them.

1 Introduction

Non-projectivity of syntactic dependency trees belongs to research topics which are of interest for a relatively wide spectrum of scholars. From the theoretical linguistics point of view, non-projectivity opens many questions related to the structure of natural language (e.g. Hajičová et al., 2004; Kuhlman and Nivre, 2006; Miletic and Urieli, 2017), while in the area of natural language processing it is relevant with respect to parsing (e.g. Gómez-Rodríguez and Nivre, 2013). In addition, non-projectivity can be understood as a violation of one of the dominant rules of the dependency grammar, namely, that a "dependent must appear in a sentence immediately adjacent to its head except that the two may be separated by dependent(s) of either words. This rule is applied recursively, so that if the inserted dependent has a dependent of its own, the latter may in turn be inserted between its own head and *the head's* head" (Ninio, 2017).[1] This rule has the decisive impact on a transfer from the two-dimensional tree-structure to the linear phonetic structure and seems to be closely connected to the so-called dependency distance

[1] Strict requirements on projectivity of dependency trees appeared much earlier, see e.g. Hays (1964).

minimization, which is, in turn, related to cognitive requirements of language users (cf. Liu et al., 2017; Ninio, 2017).[2]

Although several papers on theoretical aspects of non-projective syntactic dependency trees were published in recent past (see e.g. Ferrer-i-Cancho, 2017; Ferrer-i-Cancho et al., 2018, and references therein), it seems that no empirical study was dedicated to properties of sentences which, according to the dependency syntax formalism, are represented by non-projective trees. For a better understanding of the phenomenon of non-projectivity, it would be useful to compare properties of projective and non-projective sentences, and to investigate their relations to properties of other language units. In this paper, which can be considered a pilot study in this area, we therefore focus on the comparison of two basic aspects.

First, sentence length (throughout the paper, measured in the number of words which the sentence consists of) in these two groups will be compared. Theoretical considerations without an empirical analysis could lead to ambiguous conclusions here. On the one hand, non-projective trees could appear more often as representations of longer sentences, because longer sentences offer more possibilities to “play” with word order, and, consequently, to display this property. On the other hand, both an increasing sentence length and the appearance of non-projectivity increase the cognitive processing difficulty of a sentence, so one cannot a priori exclude the possibility that the two phenomena could compete, and that, as a result of their competition, length of non-projective sentences would not be allowed to increase too much (although, obviously, such sentences must contain at least three words). This apparent dilemma was, however, solved already. The chance that a crossing appears in a sentence (i.e., that the sentence is non-projective) increases with the increasing mean dependency distance in the sentence (Jiang and Liu, 2015; Ferrer-i-Cancho and Gómez-Rodríguez, 2016). Next, the mean dependency length tends to increase with the increasing sentence length (Ferrer-i-Cancho and Liu, 2014; Jiang and Liu, 2015). It follows that the longer sentence, the more likely it is non-projective (this hypothesis has been corroborated also empirically by Ferrer-i-Cancho et al., 2018). Ferrer-i-Cancho (2017) provides another indirect support – in random trees, the number of crossings increases with the growing number of vertices (i.e. if words in a sentence were ordered randomly, longer sentences would, again, have a higher chance to be non-projective).

Second, we will compare the frequency distributions of sentence lengths from both groups. The question is whether the same probability distribution can serve as a model in both cases; and, if the answer is positive, whether parameters of the distribution can distinguish the two groups. It can be expected that, for projective sentences, we will be able to fit the data by a special case of the very general model derived by Wimmer and Altmann (2005); in addition to being general and thus fitting well most of linguistic data, the model has also its linguistic background and its parameters are interpretable in terms of the Zipfian equilibrium of requirements of “speaker” and ”hearer” (cf. Zipf, 1949). In addition, Best (2005) already suggested some of its special cases specifically as models for sentence length, one of them for sentence length measured in the number of words. As non-projective trees can be, in a way, considered an anomaly, the model for frequency distribution of their length is much more questionable.

2 Language material and methodology

For the analysis, the Czech-PDT UD treebank is used. This treebank is based on the Prague Dependency Treebank 3.0 (Bejček et al., 2013), it consists of Czech journalistic texts from 1990s. Specifically, we used a training file named cs_pdt-ud-train.conllu from the Lindat Clarin repository (https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2895). Headings, titles, indication of a place where an article was written etc., i.e. all units which do not, in fact, represent a sentence, were removed. All units of this kind share one common property, namely, the absence of punctuation (full stop, question mark, exclamation mark), in Czech-PDT UD annotation schema. Consequently, we used this feature of the annotation to identify them.

In the study, 35,213 sentences were analyzed in total. First, we determined non-projective trees as follows. In each tree, we need to find out whether there is a word whose children's edges are crossed by its parent’s edge. For illustration, consider sentence (1)

[2] One arrives at the same conclusion - that language users prefer shorter dependency distances and thus avoid non-projective sentences - if one starts with the cognitive requirements and takes into account the least effort principle (cf. Zipf, 1949), i.e. without specific assumptions on grammar (Ferrer-i-Cancho, 2016; Ferrer-i-Cancho and Gómez-Rodríguez, 2016; Gómez-Rodríguez and Ferrer-i-Cancho, 2017).

(1) *Do* *Prahy* *měl* *přijet* *ráno*
to Prague be supposed$_{\text{PRET 3 SG.}}$ to come morning

‘*He was supposed to come to Prague in the morning*’

and examine its words one by one.

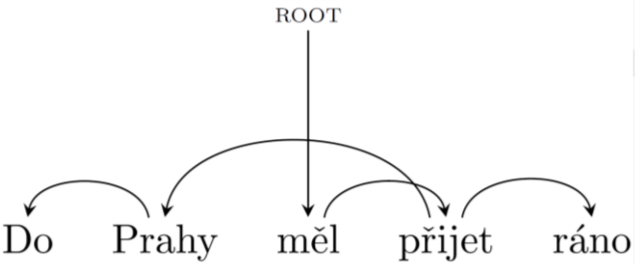

Figure 1. Syntactic relations between words in sentence (1) based on the Universal Dependencies annotation scheme. The root of the sentence is the word “*měl*”.

For each word, we look at the list of its children: does the list form an uninterrupted sequence within the sentence? If yes, the sentence is projective; if not, is the interruption caused only by the word itself (i.e., by the parent of the children under consideration)? If yes, the sentence is projective; if not, i.e. if there is also at least one other word which splits the sequence of the children, the sentence is non-projective. In sentence (1), the root, i.e. the word “*měl*”, has only one child, namely, the word “*přijet*”. Then, the word “*přijet*” has children “*Prahy*”, and “*ráno*”. The sequence of these words is interrupted not only by their parent word “*přijet*”, but also by the word “*měl*”. Thus, the sentence is non-projective (see Figure 1). This algorithm can by described by the following pseudocode (Word stands for the examined word, ID for its index, AllChildren is a zero-based sequential list of its children):

```
Word.Projectivity ← IsProjective;
 d ← 0;
 for i ← 1 to Word.AllChildren.Count - 1 do
   if (Word.AllChildren[0].ID + i + d ≠ Word.AllChildren[i].ID) then
     if (Word.ID = Word.AllChildren[0].ID + i + d) then
       Increment(d)
     else
         Word.Projectivity ← IsNonProjective;
```

The source code that was used can be found at http://milicka.cz/kestazeni/nonprojective1.zip, the function TWord.IsProjective is placed in the UDParser unit.

3 Results

Before we present results on sentence length, we shortly address the issue of proportions of non-projective trees in the treebank used. According to our analysis, non-projective trees form 8.04% of the sample. This proportion is smaller than findings presented by Havelka (2007, p. 614, Table 1) who reported 23.15% of non-projective trees in the Prague Dependency Treebank. Given that Havelka (2007) and this paper use the same treebank, but that the former study used the PDT annotation scheme while we use the UD annotation, the difference in results seems to be a consequence of using different annotation schemes. This topic deserves a more detailed study (e.g. comparing sentences which are projective according to one annotation scheme but non-projective according to the other).[3]

[3] Havelka (2007, p. 614, Table 1) found the following proportions of non-projective trees: 11.16% in Arabic (out of the total of 1,460 trees), 5.38% in Bulgarian (12,823 trees), 23.15% in Czech (72,703 trees), 15.63% in Danish (5,190 trees),

Our results on sentence length confirm the findings presented earlier by Ferrer-i-Cancho et al. (2018). Non-projective sentences in the Czech treebank we used are significantly longer (with the 95% confidence interval for the mean being $\langle 21.33; 21.93\rangle$) than projective ones (the 95% confidence interval for the mean is $\langle 16.04; 16.19\rangle$), see also Figure 2.[4]

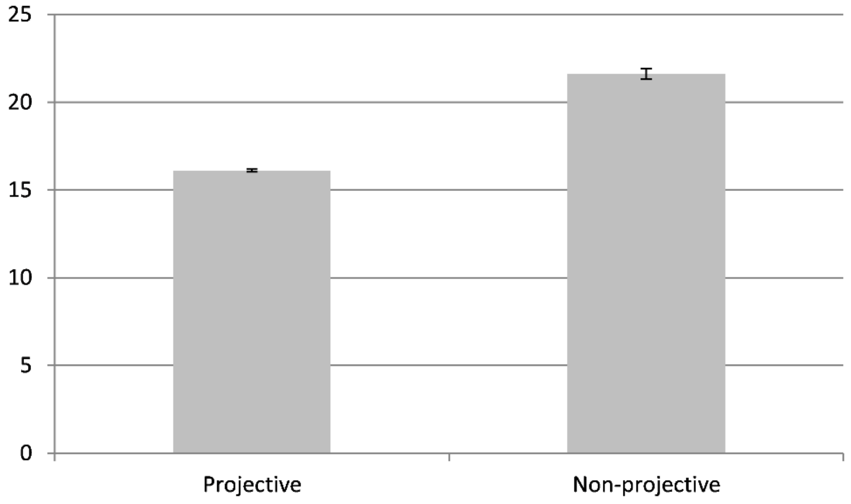

Figure 2. Length of projective and non-projective sentences in the Czech treebank (with 95% confidence intervals).

Basic descriptive statistics (which allow to formulate some – admittedly very tentative – conjectures on the comparison of lengths of projective and non-projective sentences) can be found in Table 1.

	projective	non-projective
mean	16.25	21.52
standard deviation	8.46	10.16
skewness	1.01	1.40
relative entropy	0.80	0.80

Table 1. Basic statistics on length of projective and non-projective sentences in the Czech treebank.

It is interesting that while lengths of non-projective sentences seem to be more dispersed (they achieve a higher standard deviation) and their frequency distribution more skewed, they do not differ from pro-

36.44% in Dutch (13,349 trees), 27.75% in German (39,216 trees), 5.29% in Japanese (17,044 trees), 18.94% in Portuguese (9,071 trees), 22.16% in Slovene (1.534 trees), 1.72% in Spanish (3,306 trees), 9.77% in Swedish (11,042 trees), and 11.60% in Turkish (4,997 trees). They vary quite a lot, and especially the difference between the proportions in languages so similar as Portuguese and Spanish is striking. Although we focus on the Czech treebank in this paper, we ran preliminary analyses on the proportions of non-projective sentences in several other languages using the UD annotation, with results as follows: 1.90% in Arabic (out of the total of 999 sentences), 8.04% in Czech (35,213 sentences), 0.23% in Polish (13,748 sentences), 4.81% in Russian (48,176 sentences), and 1.80% in Slovak (7817 sentences). The proportions are much lower both in Arabic and in Czech (i.e. in the two languages for which we can directly compare our results with the ones by Havelka, 2007). In addition to different annotation schemes, the differences can be caused also by the treatment of the treebanks ("[w]e take the data as is", Havelka, 2007, p. 612, vs. our approach described in Section 2 – we removed headings, titles etc., and analyzed only proper sentences). Yet another possible source of differences cannot be neglected, namely, the sentences themselves and the text which they form. The influence of text type/genre (e.g., written vs. spoken language; or, within written texts, e.g. belletristic prose, journalistic texts, scientific papers, etc.) and author on dependency syntax (in general, including non-projectivity) is a topic which, although touched in several papers (Hollingsworth, 2012; Wang and Liu, 2017; Yan and Liu, 2017; Mehler et al., 2018; Wang and Yan, 2018), is waiting for a systematic analysis.

[4] Non-projective sentences are significantly longer also in Arabic, Polish, Russian, and Slovak treebanks (cf. a short discussion on proportions of non-projective sentences at the beginning of Section 3). All these treebanks were processed in the same way as the Czech one, i.e. only proper sentences (as opposed to titles, headings etc.) were taken into account.

jective ones with respect to their relative entropies. Again, the question whether these observations represent a general tendency or whether they are specific for the Czech language (or even for this particular dependency syntax formalism) can be answered only after a more comprehensive analyses of this and related phenomena.

Best (2005) claims that frequencies of sentence lengths measured in words can be modelled by the hyperpascal distribution (cf. Wimmer and Altmann, 1999, pp. 279-281), with

$$P_x = \frac{\binom{k+x-1-s}{x-s}}{\binom{m+x-1-s}{x-s}} q^{x-s} P_0,$$

where $x = s, s+1, s+2, \ldots$ are sentence lengths; s is the shift of the distribution (in this context, the length of the shortest sentence observed); k, m, and q are free parameters.[5] He also provides a theoretical substantiation of the model. Frequencies of lengths of sentences represented by projective and non-projective trees were fitted by this distribution; however, extreme outliers (the two longest projective sentences, consisting of 78 and 162 words, and the longest non-projective sentence, with 119 words) were removed[6] before the numerical procedures for the fit were performed. After the removal of the outliers, the longest sentence consists of 76 words in case of projective sentences and 91 words in case of non-projective ones.[7]

The results of the fit are presented in Figure 3 and Table 2. Full data (also for Arabic, Polish, Russian, and Slovak) can be found at http://www.cechradek.cz/data/2019_Macutek_etal._Nonprojectivity_Length_proportions.zip .

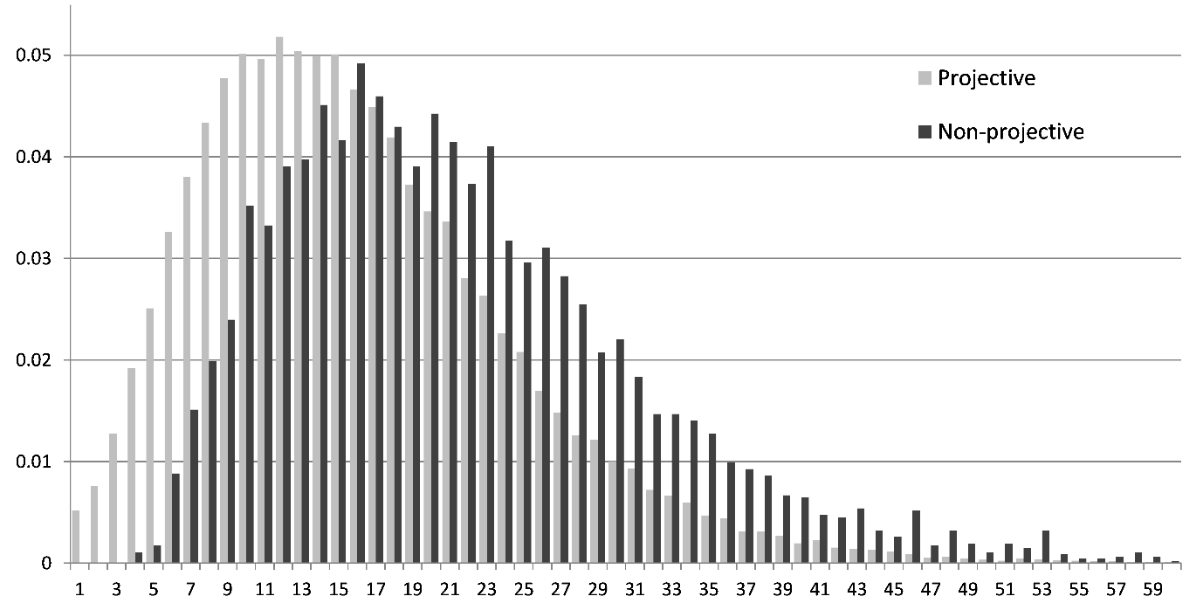

Figure 3. Relative frequencies of lengths of projective (black) and non-projective (grey) sentences in the Czech treebank.

[5] The hyperpascal distribution has three free parameters - k, m, and q. The value of P_0 is uniquely determined by the other parameters (cf. Wimmer and Altmann, 1999, p. 280).

[6] The usual boxplot-based rules for detection of outliers (i.e., outliers are values below $q_1 - 1.5IQR$ or above $q_3 + 1.5IQR$, with q_1 and q_3 being the first and the third quartile, respectively; $IQR = q_1 - q_3$ is the interquartile range, cf. Tukey, 1977) indicate too many outliers for highly skewed distributions such as ours. While there are more sophisticated versions of boxplot available for such data (e.g. the one suggested by Bruffaerts et al., 2014), in this paper we use, as a rule of thumb, a boxplot with much wider whiskers defined by $q_1 - 5IQR$ and $q_3 + 5IQR$.

[7] We remind that lengths of projective and non-projective sentences were compared (see Table 1 and Figure 2) using all sentences, i.e. prior to the removal of the outliers.

It is well-know that, in terms of the p-value, the chi-square goodness of fit test rejects practically all null hypotheses if the sample size is large enough.[8] In linguistics, it became standard to evaluate the goodness of fit of a model using the so-called discrepancy coefficient $C = \chi^2/N$, where χ^2 is the value of the test statistic in the chi-square goodness of fit test,[9] and N is the sample size. As a rule of thumb, $C \leq 0.02$ indicates a good fit; a "more tolerant" version of the rule accepts a good fit of a model if $C \leq 0.05$ (cf. Mačutek and Wimmer, 2013, where also other possibilities how to avoid the problem of large samples are mentioned). Parameters were estimated by the minimum χ^2 method (cf. Hsiao, 2006).

	projective	non-projective
k	9.14	1.66
m	3.84	0.20
q	0.74	0.87
s	1	5
N	32379	2831
C	0.0073	0.0384

Table 2. Fitting the hyperpascal distribution to frequency distribution of length of projective and non-projective sentences in the Czech treebank.

Values of parameters k and m for projective and non-projective sentences are quite far from each other (we postpone testing and attempts to interpret both the parameter values and their differences until data from more languages are available). It means that the two frequency distributions differ in their shape, not only in the shift to the right represented by the increase of parameter s. The relatively worse (but still acceptable) fit of the hyperpascal distribution to length frequencies of non-projective sentences can be explained by their smaller number, and perhaps also by the fact that they can be considered, in a way, an anomaly, and it cannot be a priori excluded that their properties (among them their length) can differ from the "normal" (i.e. projective) sentences.

4 Conclusion and perspectives

Our results provide a further empirical corroboration of the hypothesis that non-projective sentences are longer than projective ones. Moreover, we show that frequency distribution of sentence length can be fitted by the same model in the two groups, albeit with different parameter values.

In addition to results, the paper also opens several questions. First, proportions of non-projective sentences vary not only across languages, but they depend also on the annotation scheme (such as PDT or UD), and probably on genre and author of a text as well. A systematic study, e.g. one where three out of the four "variables" under consideration (i.e., language, annotation scheme, genre, author) are fixed and the influence of the fourth one is investigated, is necessary before this problem can be at least partially solved.

Second, while word is one of reasonable units in which sentence length can be measured, it is not the only one possible – on the contrary, quantitative approaches to language modelling prefer immediate "neighbours" in the hierarchy of language units (cf. e.g. Köhler, 2012). Sentence length measured in the number of clauses could reveal other properties of non-projective sentences (and their differences from projective ones).

Third, as we suppose that "[n]o property of things or linguistic entities is isolated; each of them is in at least one relation to the other properties of the same thing, or those of other things" (Altmann, 1993; cf. also Köhler, 2005, who tries to build a general language theory which encompasses different language units, their properties and their interrelations and mutual influences), neither is sentence length. The Menzerath-Altmann law (in general cf. Cramer, 2005) predicts that longer sentences should consist of shorter clauses (cf. Köhler, 1982; Heups, 1983, Teupenhayn and Altmann, 1984; the law seems to be

[8] Browne and Cudeck (1993) wrote that "… goodness-of-fit tests are often more a reflection on the size of the sample than on the adequacy of the model". This problem is not specific to goodness-of-fit tests only, one encounters it whenever a statistical test with a fixed level of significance is used (cf. e.g. Kunte and Gore, 1992).

[9] $\chi^2 = \sum_{i=s}^{L} \frac{(f_i - NP_i)^2}{NP_i}$, where f_i is the observed frequency of sentences with length i, NP_i is the frequency of sentences with length i predicted by the model, and L is the length of the longest sentence observed.

valid also within the dependency syntax formalism - see Mačutek et al., 2017, where results on the relation between lengths of clauses and phrases, i.e. one level lower, are presented). The question is whether the non-projective sentences “obey” this law; if yes, whether the parameters in the mathematical formulation of the law reflect the difference between them and projective sentences (we allow ourselves to formulate the hypothesis that the decrease of clause length for longer non-projective sentences will be steeper, which could compensate for their higher cognitive processing difficulty).

Acknowledgements

Supported by research projects VEGA 2/0054/18 (J. Mačutek) and ERDF CZ.02.1.01/0.0/0.0/16_019/0000734 “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (J. Milička).

References

Gabriel Altmann. 1993. Science and linguistics. In Reinhard Köhler and Burghard B. Rieger (eds.), *Contributions to Quantitative Linguistics*, pp. 3-10. Kluwer, Dordrecht.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Karl-Heinz Best. 2005. Satzlänge. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 298-304. de Gruyter, Berlin / New York.

Michael W. Browne and Robert Cudeck. 1993. Alternative ways of assessing model fit. In Kenneth A.Bollen and J.Scott Long (eds.), *Testing Structural Equation Models*, pp. 136-161. SAGE, Newbury Park (CA).

Christopher Bruffaerts, Vincenzo Verardi, and Catherine Vermandele. 2014. A generalized boxplot for skewed and heavy-tailed distributions. *Statistics and Probability Letters*, 95:110-117.

Irene M. Cramer. 2005. Das Menzeratsche Gesetz. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 659-688. de Gruyter, Berlin / New York.

Ramon Ferrer-i-Cancho. 2016. Non-crossing dependencies: Least effort, not grammar. In Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Frank-Job (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pp. 203-234. Springer, Berlin / Heidelberg.

Ramon Ferrer-i-Cancho. 2017. Random crossings in dependency trees. *Glottometrics*, 37:1-12.

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

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2018. Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, 493:311-329.

Ramon Ferrer-i-Cancho and Haitao Liu. 2014. The risks of mixing dependency lengths from sentences of different length. *Glottotheory*, 5(2):143-155.

Carlos Gómez-Rodríguez and Ramon Ferrer-i-Cancho. (2017). Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96:062304.

Carlos Gómez-Rodríguez and Joakim Nivre. 2013. Divisible transition systems and multiplanar dependency parsing. *Computational Linguistics*, 39(4):799-845.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81:5-22.

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

David G. Hays. 1964. Dependency theory: A formalism and some observations. *Language*, 40(4):511-525.

Gabriela Heups. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler and Joachim Boy (eds.), *Glottometrika 5*, pp. 113-133. Brockmeyer, Bochum.

Charles Hollingsworth. 2012. Using dependency-based annotations for authorship identification. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (eds.), *Text, Speech, and Dialogue*, pp. 314-319. Springer, Cham.

Cheng Hsiao. 2006. Minimum chi-square. In Samuel Kotz and Norman L. Johnson (eds.), *Encyclopedia of Statistical Sciences, Vol. 7*, pp. 4812-4817. Wiley, Hoboken (NJ).

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications – based on a parallel English-Chinese treebank. *Language Sciences*, 50:93-104.

Reinhard Köhler. 1982. Das Menzerathsche Gesetz auf Satzebene. In Werner Lehfeldt and Udo Strauss (eds.), *Glottometrika 4*, pp. 103-113. Brockmeyer, Bochum.

Reinhard Köhler. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 760-775. de Gruyter, Berlin / New York.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter, Berlin / Boston.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the COLING/ACL 2006*, pp. 507-514. ACL.

Sudhakar Kunte and Anil P. Gore. 1992. The paradox of large samples. *Current Science*, 62:393-395.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171-193.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni and Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 100-107. Linköping University Electronic Press, Linköping.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3):227-240.

Alexander Mehler, Wahed Hemati, Tolga Uslu, and Andy Lücking. 2018. A multidimensional model of syntactic dependency trees for authorship attribution. In Jingyang Jiang and Haitao Liu (eds.), *Quantitative Analysis of Dependency Structures*, pp. 315-347. de Gruyter, Berlin / Boston.

Aleksandra Miletic and Assaf Urieli. 2017. Non-projectivity in Serbian: Analysis of formal and linguistic properties. In Simonetta Montemagni and Joakim Nivre (eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 135-144. Linköping University Electronic Press, Linköping.

Anat Ninio. 2017. Projectivity is the mathematical code of syntax. Comment on "Dependency distance: A new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews*, 21:215-217.

Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath's law. In Joachim Boy and Reinhard Köhler (eds.), *Glottometrika 6*, pp. 127-138. Brockmeyer, Bochum.

John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading (MA).

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135-147.

Yaqin Wang and Jianwei Yan. 2018. A quantitative analysis on literary genre *essay*'s syntactic features. In Jingyang Jiang and Haitao Liu (eds.), *Quantitative Analysis of Dependency Structures*, pp. 295-314. de Gruyter, Berlin / Boston.

Gejza Wimmer and Gabriel Altmann. 1999. *Thesaurus of Univariate Discrete Probability Distributions*. Stamm, Essen.

Gejza Wimmer and Gabriel Altmann. 2005. Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pp. 791-807. de Gruyter, Berlin / New York.

Jianwei Yan and Siqi Liu. 2017. The distribution of dependency relations in *Great Expectations* and *Jane Eyre*. *Glottometrics*, 37:13-33.

George K. Zipf. 1949. *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology.* Addison-Wesley, Cambridge (MA).