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

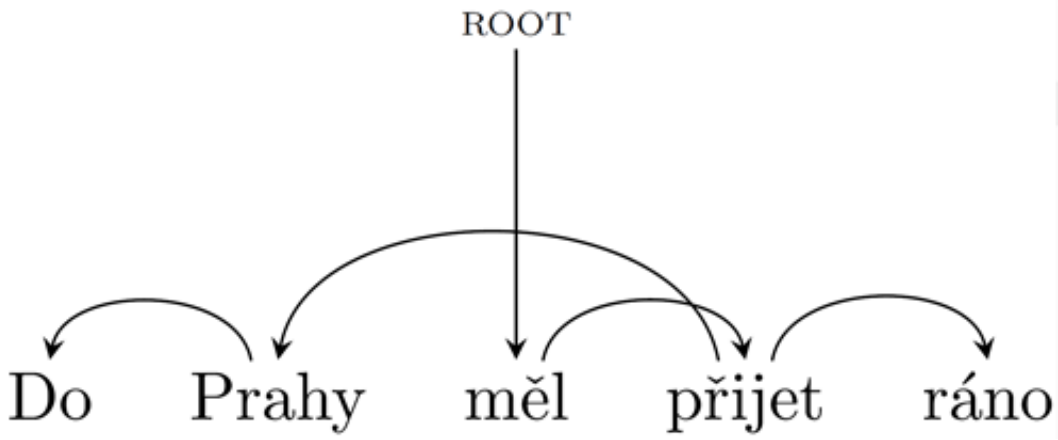
Ján Mačutek (Comenius University in Bratislava, Slovakia)
Radek Čech (University of Ostrava, Czech Republic)
Jiří Milička (Charles University in Prague, Czech Republic)

Non-projective sentence

Do Prahy měl přijet ráno.

He was supposed to come to Prague in the morning.

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



Non-projectivity is “non-standard” (1/2)

- a “violation” of one of the dominant rules of the DG
  - 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” (A. 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)

Non-projectivity is “non-standard” (2/2)

- cognitive requirements
  - language users prefer shorter dependency distances and thus avoid non-projective sentences – a result based cognitive requirements and the Zipfian least effort principle, i.e. without specific assumptions on grammar (e.g. R. Ferrer-i-Cancho, 2016, Non-crossing dependencies: Least effort, not grammar. In A. Mehler et al. (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pp. 203-234, Springer, Berlin / Heidelberg)

Non-projectivity & sentence length (1/4)

Are non-projective sentences longer or shorter than projective ones?

Methodological aspects:

- sentence length measured in the number of words the sentence contains
- not the usual approach in the QL framework – length of a language unit is usually measured in the number of its “direct neighbours” in the hierarchy of units (e.g. words in morphemes or syllables, clauses in words, sentences in clauses, …)
- our choice is motivated by technical reasons (much easier to get the number of words in a sentence than a number of clauses)

Non-projectivity & sentence length (2/4)

Are non-projective sentences longer or shorter than projective ones?

A speculative (i.e. non-empirical) considerations lead to two different answers.

Non-projectivity & sentence length (3/4)

- non-projective trees in **longer** sentences than projective trees
  - more „space“ for the realization of non-projectivity
  - random models

- non-projective trees in **shorter** sentences than projective trees
  - cognitive processing difficulty (both an increasing sentence length and the appearance of non-projectivity make a sentence more difficult to process)
  - theoretically, language users could “forbid” making long sentences (which are already difficult to process because of their length) even more complicated (by introducing non-projectivity)

Non-projectivity & sentence length (4/4)

Good news – the dilemma is solved!

- the chance that a sentence is non-projective increases with the increasing mean dependency distance (R. Ferrer-i-Cancho, C. Gómez-Rodríguez, 2016, Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320-328)
- the mean dependency length tends to increase with the increasing sentence length (J. Jiang, H. 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)
- it follows that the longer the sentence, the more likely it is non-projective (corroborated also empirically by R. Ferrer-i-Cancho et al., 2018, Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, 493:311-329)

Frequency distribution of sentence lengths

- the same or different **models**?
- if the same model, the same or different **parameters**?

Language material and methodology

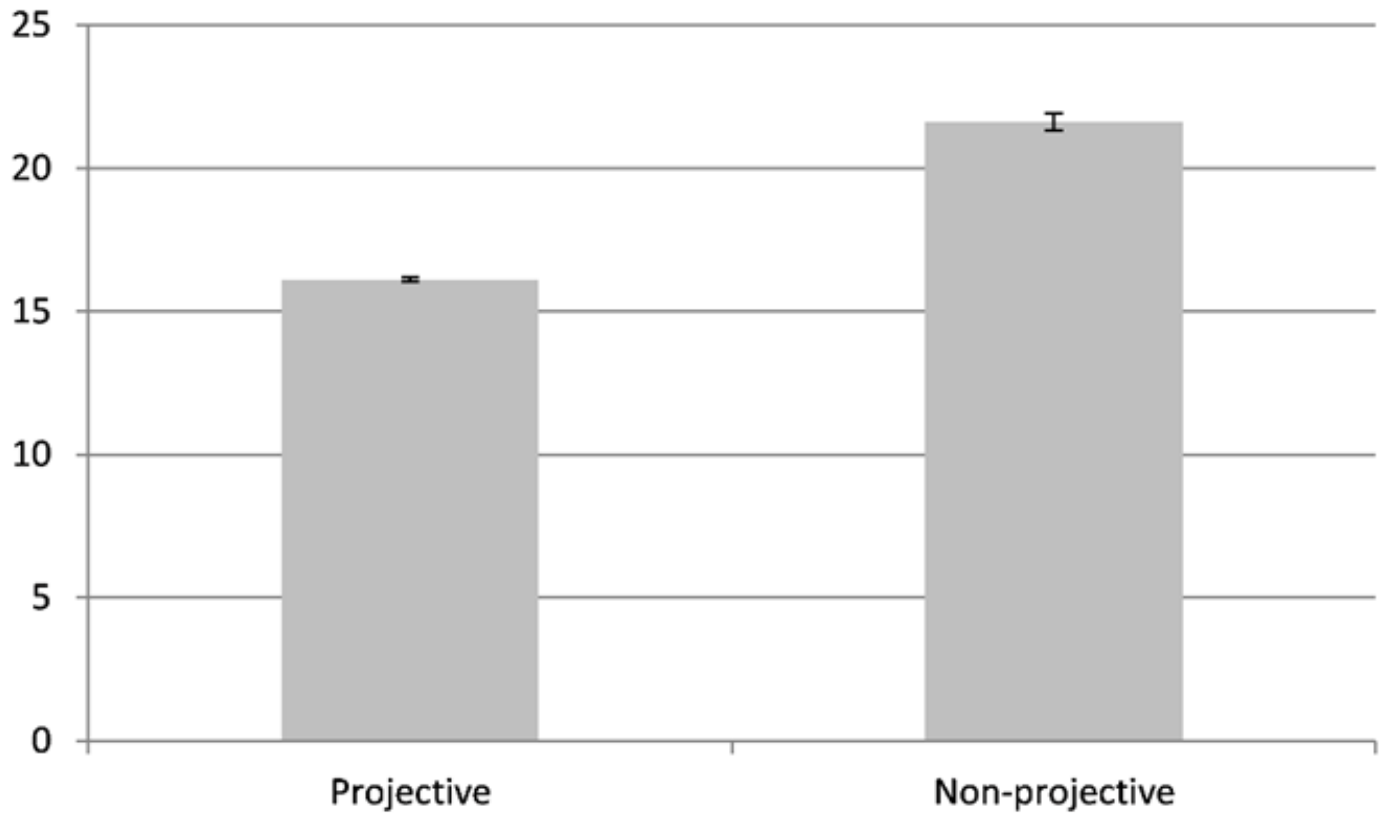
- Czech-PDT UD treebank (based on PDT 3.0)
- 35,213 sentences
- UD annotation scheme

- proportion of non-projective trees in the sample is 8.04%
  - Havelka (2007) detects 23.15% in the PDT

    J. 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

Results

	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

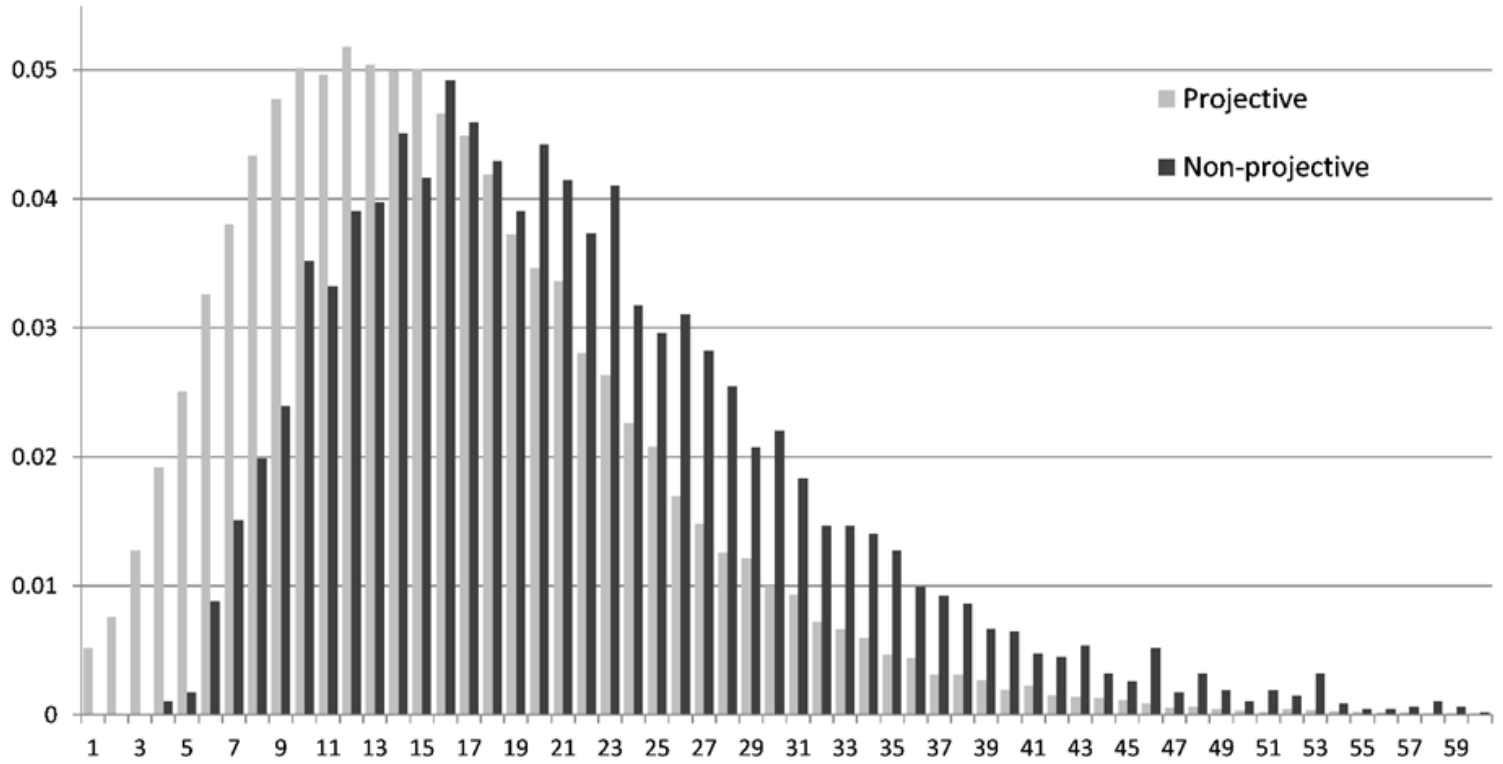


Results

- the hyper-Pascal distribution (suggested by K.-H. Best, 2005, Satzlänge, in R. Köhler et al. (eds.), *Quantitative Linguistics. An International Handbook*, pp. 298-304. de Gruyter, Berlin / New York)

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

	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



Conclusions & perspectives

- non-projective sentences are longer than projective ones
- frequency distribution of sentence length
  - the same model (a special case of a very general frequency distribution basen on the Zipfian assumptions on the equilibrium between the “forces” of the speaker and the hearer)
  - different parameter values

- open questions
  - impact of the annotation scheme
  - impact of language, genre, author,...
  - relations to other properties

Merci pour votre attention!