CLIQS: Crosslinguistic Investigations in Quantitative Syntax

Richard Futrell [(MIT)], Kyle Mahowald [(MIT)], and Edward Gibson [(MIT)]
Contact: futrell@mit.edu

Quantitative Syntax

Quantitative Properties of Languages

Quantitative properties of language such as Zipf's Law (Zipf, 1936) have attracted much attention. But the known quantitative universals are only about what can be easily calculated given masses of unannotated text: i.e. mostly frequency distributions. **But what about quantitative properties of *syntax?***

Here we present some results from investigations of:

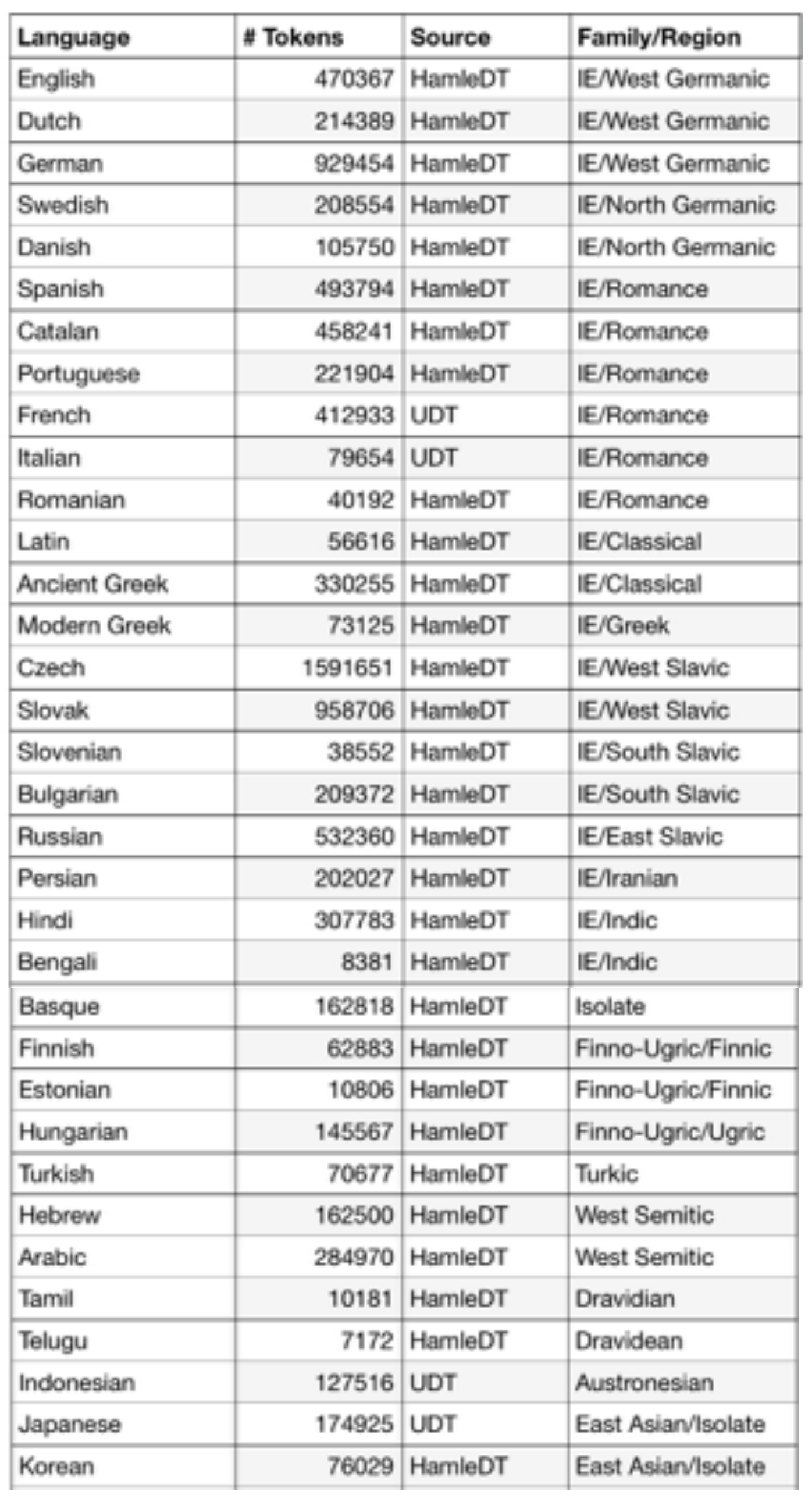
- **Word Order Freedom**: Languages that allow many word orders in principle might vary in how much freedom they really exhibit. Word order variability is supposed to correlate with the presence of case marking (e.g., Sapir, 1923; Kiparsky, 1997; McFadden, 2003). If we want to know if *more* variability implies *more* case marking, we need quantitative measures of word order freedom.
- **Dependency Length**: A large body of research (e.g. Hawkins, 1994; Hawkins, 2004; Gildea & Temperley, 2010; Tily, 2012) argues that, for processing reasons, languages should evolve to minimize the linear length between heads and their dependents. Average dependency length is a quantitative property of language syntax which should be minimized.

Data Sources

Recent interest in multilingual **dependency parsing** in NLP has resulted in the release of dependency-parsed corpora in many languages (e.g. the CoNLL 2007 Shared Task (Nivre et al., 2007)).

Differences in annotation have been harmonized by two separate projects: **HamleDT** (Zeman et al. 2012) and the **Google Universal Dependency Treebank** (MacDonald et al., 2013). We have combined these corpora and done further harmonization.

The corpora are mostly newspaper text and novels. Exceptions are the Japanese corpus, which is spoken, and Latin and Ancient Greek, which include metered poetry.



Language	# Tokens	Source	Family/Region
English	470367	HamleDT	IE/West Germanic
Dutch	214389	HamleDT	IE/West Germanic
German	929454	HamleDT	IE/West Germanic
Swedish	208554	HamleDT	IE/North Germanic
Danish	105750	HamleDT	IE/North Germanic
Spanish	493794	HamleDT	IE/Romance
Catalan	458241	HamleDT	IE/Romance
Portuguese	221904	HamleDT	IE/Romance
French	412933	UDT	IE/Romance
Italian	79654	UDT	IE/Romance
Romanian	40192	HamleDT	IE/Romance
Latin	56616	HamleDT	IE/Classical
Ancient Greek	330255	HamleDT	IE/Classical
Modern Greek	73125	HamleDT	IE/Greek
Czech	1591651	HamleDT	IE/West Slavic
Slovak	958706	HamleDT	IE/West Slavic
Slovenian	38552	HamleDT	IE/South Slavic
Bulgarian	209372	HamleDT	IE/South Slavic
Russian	532360	HamleDT	IE/East Slavic
Persian	202027	HamleDT	IE/Iranian
Hindi	307783	HamleDT	IE/Indic
Bengali	8381	HamleDT	IE/Indic
Basque	162818	HamleDT	Isolate
Finnish	62883	HamleDT	Finno-Ugric/Finnic
Estonian	10806	HamleDT	Finno-Ugric/Finnic
Hungarian	145567	HamleDT	Finno-Ugric/Ugric
Turkish	70677	HamleDT	Turkic
Hebrew	162500	HamleDT	West Semitic
Arabic	284970	HamleDT	West Semitic
Tamil	10181	HamleDT	Dravidian
Telugu	7172	HamleDT	Dravidean
Indonesian	127516	UDT	Austronesian
Japanese	174925	UDT	East Asian/Isolate
Korean	76029	HamleDT	East Asian/Isolate

Table 1. Corpora available and their properties.

Dependency Formalism

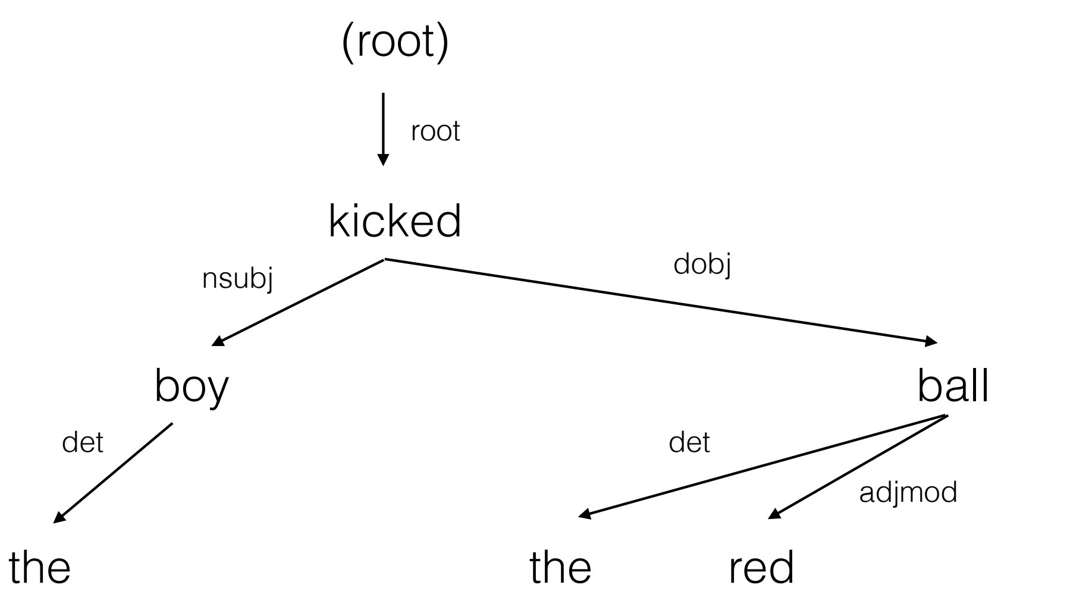


Figure 1. An example of the dependency formalism used in the corpora.

Word Order Freedom

Using Entropy to Quantify Variability

We use **conditional entropy** to quantify order variability **conditioned on relation type**. This is interpretable as the **degree of uncertainty about order within relations**.

Entropy measures are **sensitive to sample size**: To make sure this is not influencing our results, we also calculated the measures presented here on small subsets (1000 tokens) of the corpora; we found very little difference in the resulting numbers (r>.97 between the measures calculated on subsets and measures calculated on the whole corpora).

Branching Direction Entropy

The conditional entropy of head direction conditioned on relation type. (Bounded between 0 and 1, where 0 is totally deterministic, and 1 means there is total uncertainty about head direction.)

There is great variability in BDE: from 0 for Japanese and Korean to near .75 for Finnish and Estonian. No language exceeds .75.

Some related languages are very similar (e.g., Finnish and Estonian; Telugu and Tamil). Other families are more variable (e.g. Romance.)

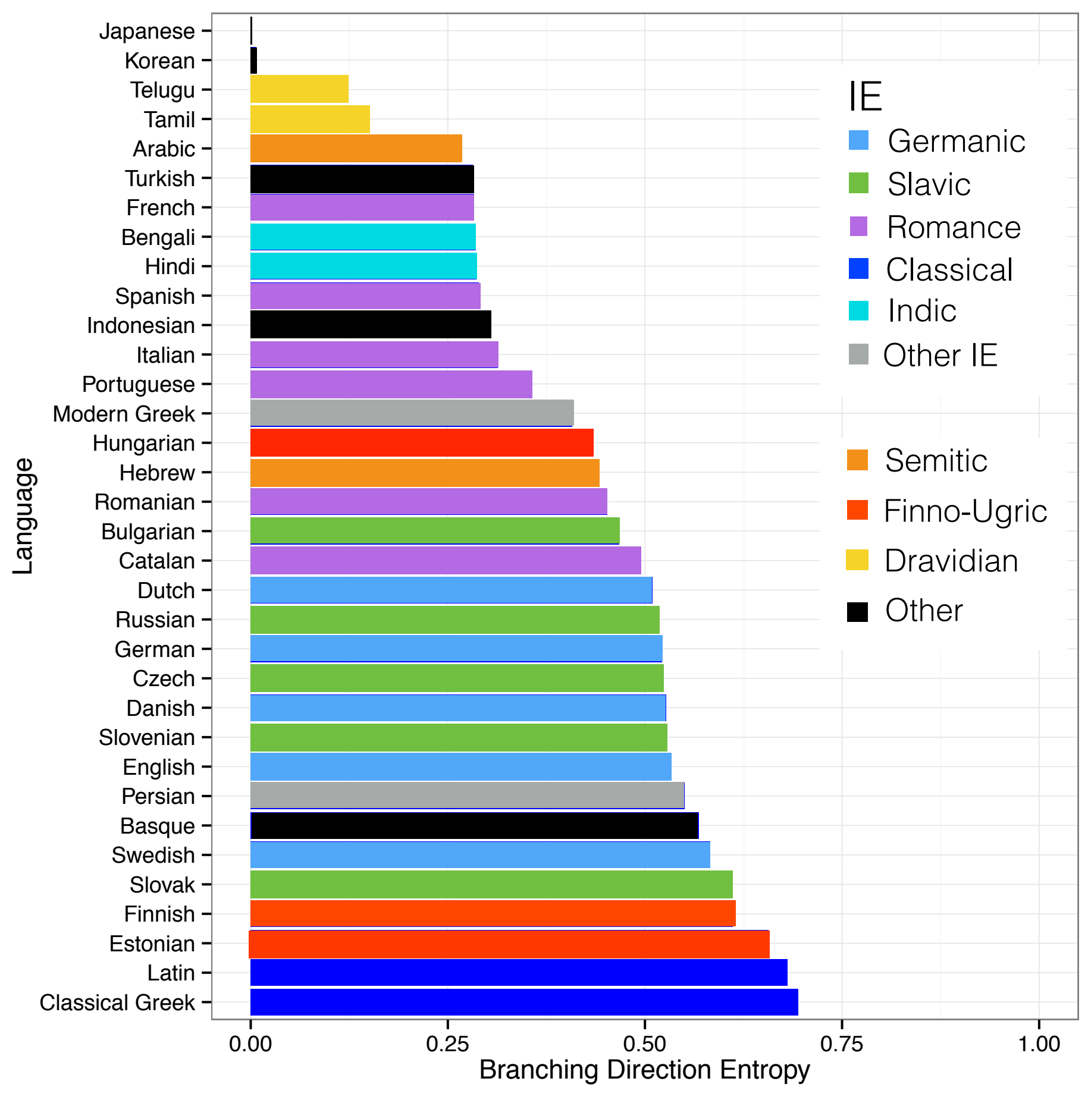


Figure 2. Overall Branching Direction Entropy for 34 languages, colored by language family/area.

Order Entropy of Subj and Obj

Here we show the entropy of the order of **nsubj** and **dobj** relations under verbs where both are present. This is the word order variability for subjects and objects.

Languages are colored for their case marking system. **High-variability languages all have case marking**, but **many case-marked languages have low variability.**

Also: all SOV and VSO languages here are case-marked, which fits a noisy-channel communication account of case marking (Gibson et al, 2012).

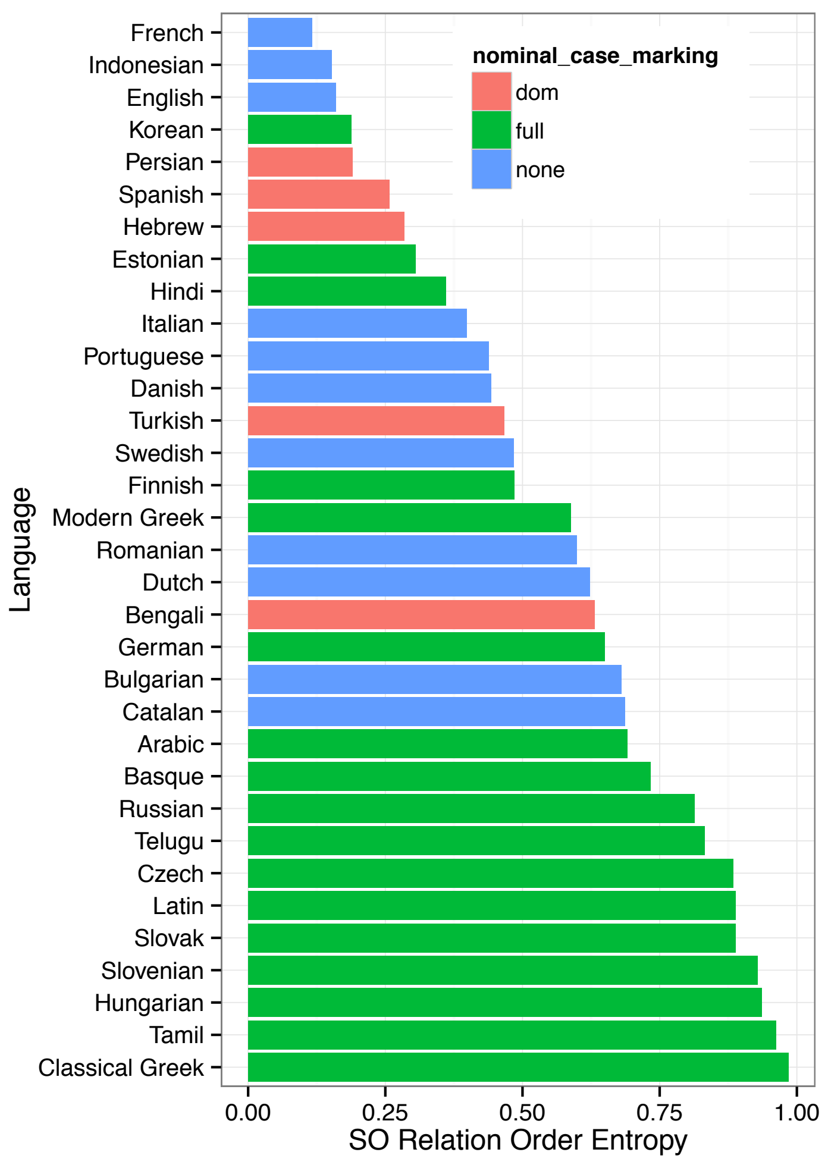


Figure 3. Order entropy of subjects and objects, colored by case marking system.

Dependency Length

Frequency Distribution of Dependency Lengths

The frequency distribution of dependency lengths has been shown to be Zipfian for a few languages (eg Chinese: Liu, 2007). Here we show what seems to be a power-law distribution for dependency lengths in all 34 languages.

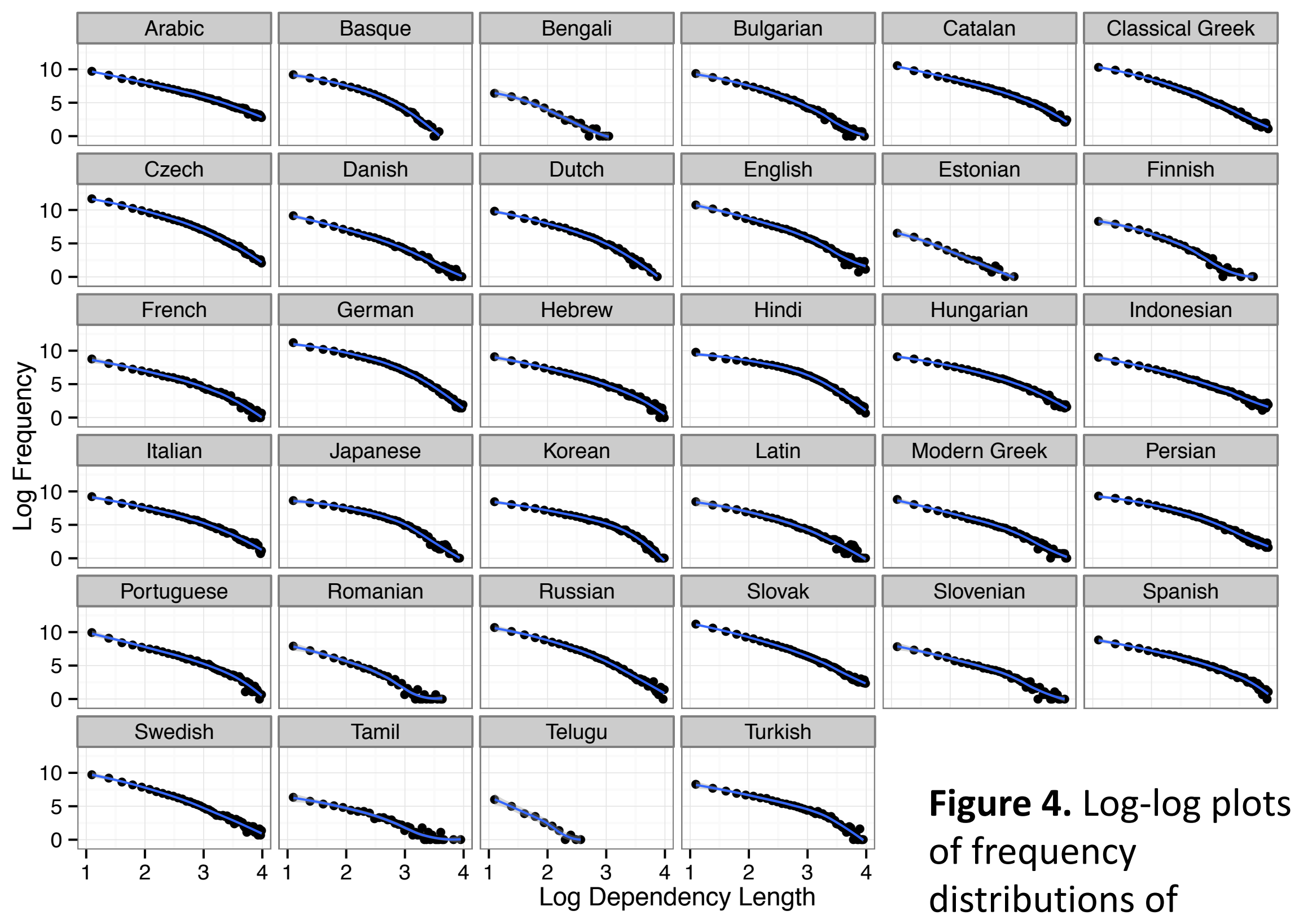


Figure 4. Log-log plots of frequency distributions of dependency lengths.

Dependency Length Minimization

Gildea & Temperley (2007; 2010) show how to calculate the **projective linearization** that **minimizes dependency length** for any dependency tree. They compare the observed dependency lengths in English and German to their minimal baseline, and also to the dependency lengths of **random projective linearizations**.

Here we find that **dependency length is minimized** in all languages of the sample, but to varying degrees.

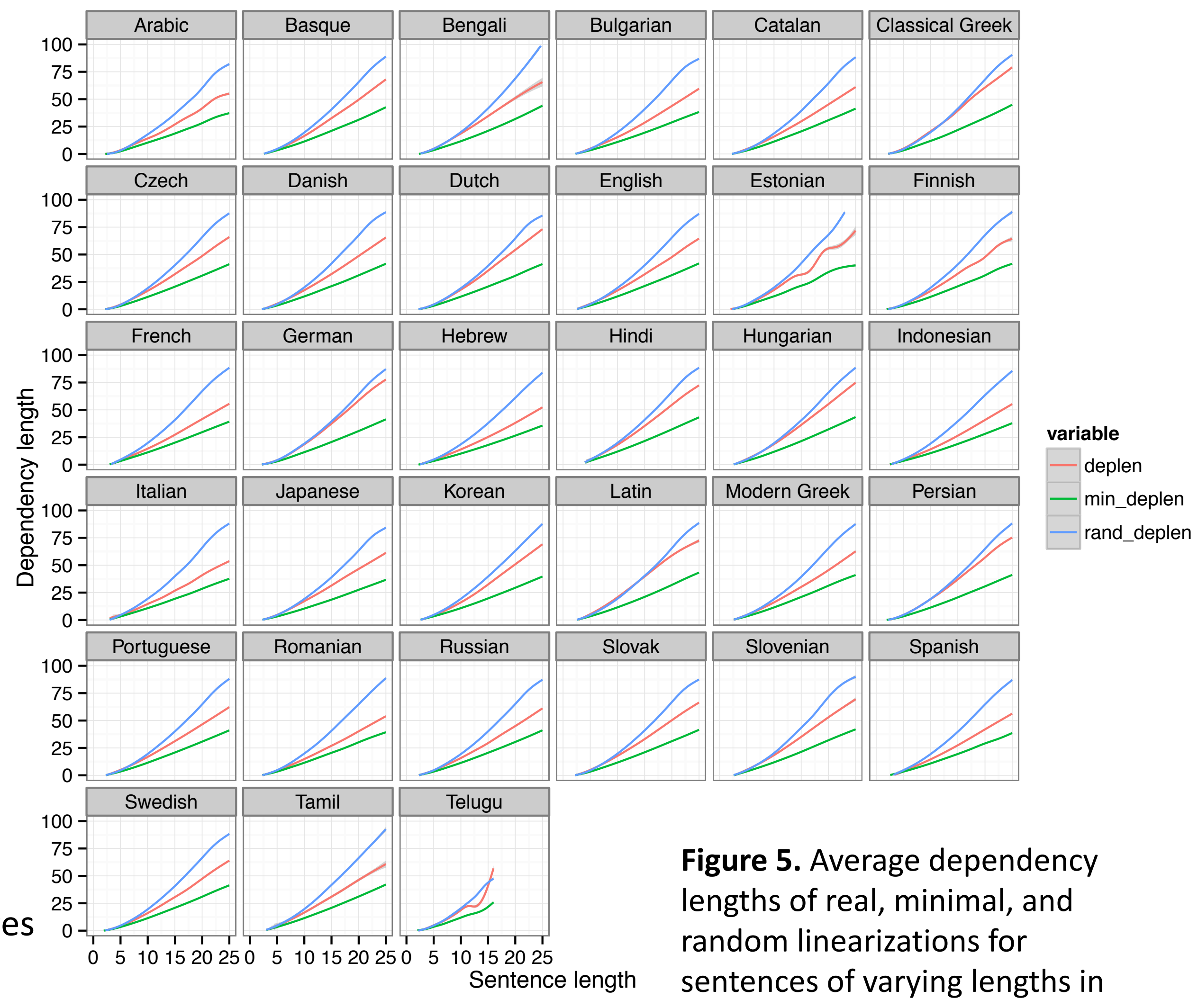


Figure 5. Average dependency lengths of real, minimal, and random linearizations for sentences of varying lengths in the corpora.

Dependency Length and Word Order Freedom

We find a **weak positive correlation** (r=.45, p=.04) between word order freedom and dependency length.

This should be unsettling to proponents of dependency length minimization: Are languages using word order freedom to increase their dependency length?

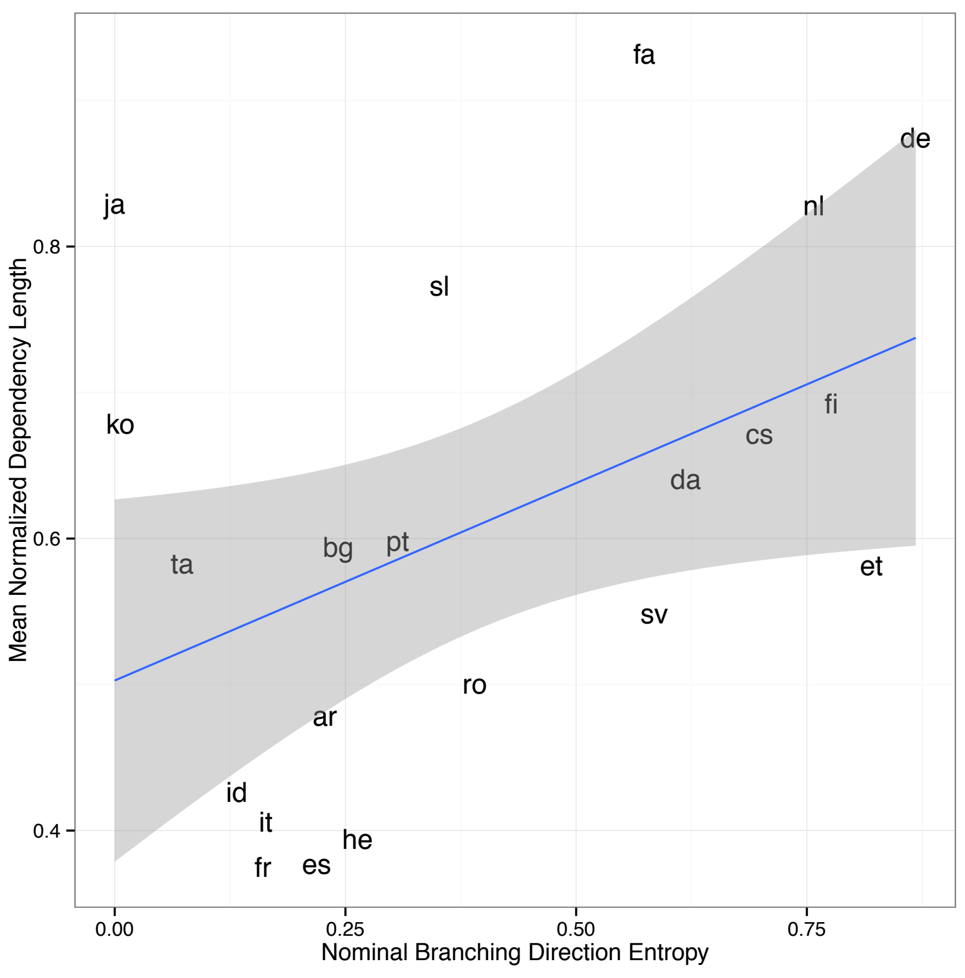


Figure 6. Branching Direction Entropy and mean normalized dependency length. Excluding Latin and Ancient Greek. Normalized dependency length: The length of an observed dependency, scaled between 0 (the length of that dependency in the minimal projective linearization) and 1 (its average length in a random projective linearization).

Conclusions

Dependency-parsed corpora make **typology of quantitative syntax** possible. We find results that are broadly consistent with previous claims about universal pressures on quantitative syntax, but with complications.

Using dependency corpora we have developed easily interpretable measures of **word order freedom** and shown that **high word order variability of subjects and objects implies case marking**, but not vice versa.

We have shown that **dependency lengths are minimized** across varied languages. But dependency length seems to correlate with word order freedom.

Besides the work presented, we believe that the measures and methods developed here can be used to quantitatively answer long-standing questions about cross-linguistic syntactic phenomena.

References

Gildea, D. & Temperley, D. (2010). Do Grammars Minimize Dependency Length? *Cognitive Science* 34: 286–310.

Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the Association for Computational Linguistics* (pp. 184–191). Prague: Association for Computational Linguistics.

Hawkins, J. (1994). *A Performance Theory of Order and Constituency.* Cambridge University Press.

Hawkins, J. (2004). *Efficiency and Complexity in Grammars.* Oxford University Press.

Kiparsky, P. (1997). The rise of positional licensing. In *Parameters of morphosyntactic change*, ed. Ans van Kemenade and Nigel Vincent, 460–494. Cambridge University Press.

Liu H. (2010). Dependency direction as a means of word-order typology: a method based on dependency treebanks. *Lingua* 120(6): 1567-1578.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics* 15: 1-12.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castello and J. Lee. (2013). Universal Dependency Annotation for Multilingual Parsing. Association of Computational Linguistics (ACL), 2013.

McFadden, T. (2003). On Morphological Case and Word-Order Freedom. In P*roceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society.*

Sapir, E. (1921). *Language, an introduction to the study of speech.* New York: Harcourt, Brace and Co.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J. (2012): HamleDT: To Parse or Not to Parse? In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC 2012), pp. 2735-2741. ELDA, İstanbul, Turkey.

Zipf, G. (1936). The Psychobiology of Language. London: Routledge.