The effect of register on dependency length in two flexible languages

Alex Kramer
(University of Michigan)

Keywords: Corpus linguistics, Japanese, Russian, constituent order, comparative

Recent corpus studies employing dependency corpora have found that SOV-dominant languages such as Japanese and Korean have longer dependency lengths and more rigid word order than might be expected given their typological profiles (Futrell et al. 2015 and Levshina 2019). These languages share features such as argument drop and flexible constituent order, which can result in shorter dependency lengths. However, speakers may be less likely to use these features in more formal registers (Biber 1993 and Wälchli 2009), and indeed, the Japanese data used in these studies came from formal written and spoken sources.

In order to investigate the effect of genre on dependency length, I constructed a corpus of informal spoken Japanese by scraping the captions from the videos of popular YouTubers. Captions were pre-processed in Python and parsed using the StanfordNLP dependency parser (Qi et al. 2018). Replicating the methods from Futrell et al. (2015), I calculated and compared the total combined length of the dependencies (a) in each observed sentence, (b) a random permutation of each sentence, and (c) an "optimized" permutation of each sentence that minimized the distances between heads and dependents.

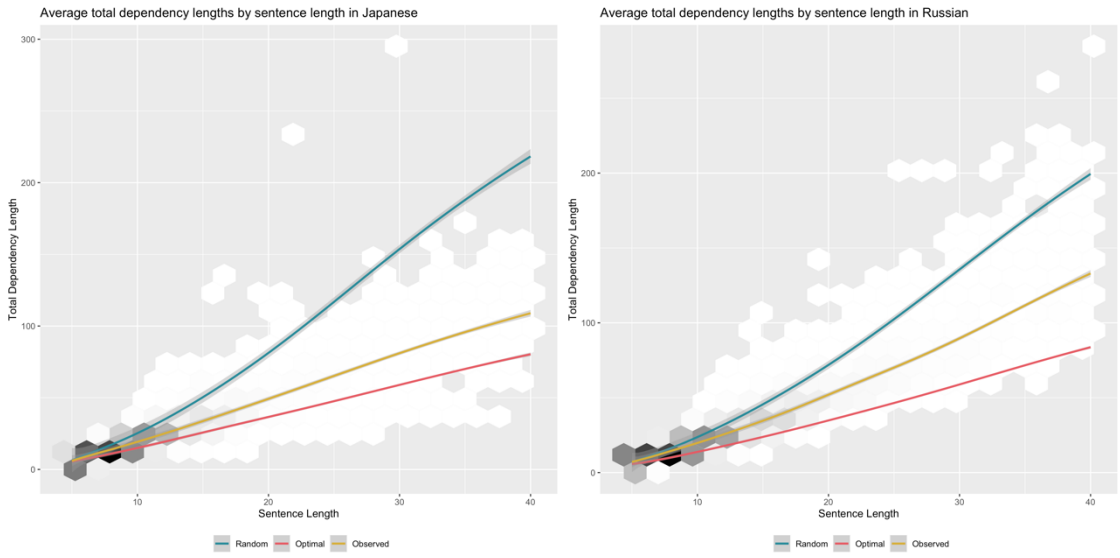
Comparison of the observed dependencies to the random and optimal baselines revealed overall shorter dependency lengths than found in Futrell et al. (2015), in which more head-final languages like Japanese, Korean, and Turkish showed less minimization than more head-initial languages. In fact, it is likely that the true Japanese dependencies are even shorter: Manual analysis found that non-canonical orders, many of which reduce dependency length, are common in the data; however, the parser mis-parses main clause verbs in these orders as modifiers, artificially lengthening dependencies.

Japanese allows flexible order and argument dropping, both of which may affect dependency lengths; argument drop, which is influenced by register and context, is extremely common in spoken Japanese (Nariyama 2000). To disentangle the effects of these two factors and extend this comparison to another flexible language, I replicated this method with Russian, a language with more restricted argument dropping. Unlike Japanese, Russian dependencies patterned similarly to the results of Futrell et al. (2015). While not a perfect comparison (e.g. Russian is canonically SVO; Japanese is SOV), these results may suggest an important role for argument dropping in dependency length minimization in spoken Japanese.

Diversifying the dependency corpora used in typological research via YouTube captions allowed for the comparison of informal spoken Japanese and Russian, which revealed different patterns of dependency length minimization across languages and registers. Future work aims to improve the parser's ability to handle non-canonical sentence orders and other aspects of informal

speech. In addition, to understand how features such as argument dropping and flexibility contribute to minimizing dependency lengths, this method should be extended to languages that vary systematically with respect to canonical order, head- versus dependent-marking, and other potentially relevant features. As computational models improve, informal sources such as YouTube have the potential to become powerful tools for uncovering typological patterns that are not present in more formal registers and written modalities.

Figures



References

Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, 19(2), 24.

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

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, *23*(3), 533–572. https://doi.org/10.1515/lingty-2019-0025

Nariyama, S. (2000). *Referent identification for ellipted arguments in Japanese* [Ph.D. Thesis, Universtiy of Melbourne]. http://minerva-access.unimelb.edu.au/handle/11343/39534

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160-170. https://doi.org/10.18653/v1/K18-2016

Wälchli, B. (2009). Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, *13*(1). https://doi.org/10.1515/LITY.2009.004