Dependency lengths in speech and writing: A cross-linguistic comparison via YouDePP, a pipeline for scraping and parsing YouTube captions

Alex Kramer
Department of Linguistics
University of Michigan, Ann Arbor
arkram@umich.edu

1 Introduction

Dependency length minimization (DLM)—the tendency for grammars to minimize the distances between dependents and their heads—has been found to be present across languages (Futrell et al., 2015a, 2020). However, while all languages have been found to minimize, they do not all appear to minimize to the same extent. In particular, these studies have found that some predominantly head-final languages—in particular, Japanese, Turkish, and Korean—seem to minimize dependencies less than other languages, especially predominantly head-initial languages such as Italian and Irish.

An open question, however, is what these minimization patterns look like across different registers, genres, and modalities. Cross-linguistic collections of dependency corpora such as Universal Dependencies version 2.6 (Zeman et al., 2020) largely represent written language. Written and formal varieties of languages, however, can differ from spoken and informal varieties in ways that may have consequences for the study of cross-linguistic phenomena such as DLM (Biber, 1992, 1993). Japanese and Russian, for example, allow both argument drop and flexible word order, with these features being more common in informal speech than in formal speech and writing (Nariyama, 2000; Ueno and Polinsky, 2009; Zdorenko, 2010).

Given these differences, it is desirable to explore ways of generating cross-linguistic corpora, and particularly spoken-language corpora, that are comparable in genre and register. Parallel speech corpora, such as the Parallel Corpus for Typology or ParTy corpus (Levshina, 2017), which consists of subtitles of popular films, are one way of achieving this aim. Subtitles are advantageous as a data source because they are freely available for download in a wide variety of languages and tend to represent a variety of the language that is closer to informal speech than many other parallel corpora (Levshina, 2017). At the same time, scripted dialogue still differs from truly spontaneous speech (Quaglio, 2008; Levshina, 2017), and as with all parallel corpora, source languages may influence target languages in translation (Johansson and Hofland, 1994).

YouTube, with its worldwide popularity and ever-growing number of videos, is an attractive source of naturalistic, and often spontaneous, speech. Many videos include captions, which may be provided by the uploader, fans of the channel, or professional captioning services, or may be generated automatically via speech-to-text[1]. Previous work has used automatically-generated captions to reveal differences in speech rates across regions of the United States (Coats, 2020); however, YouTube captions have yet to be employed in large-scale, cross-linguistic typological work.

Here, I compare dependency length growth rates between written corpora available through Universal Dependencies 2.6 and highly informal spoken corpora collected via the YouTube Dependency Parsing Pipeline, or YouDePP (Appendix A) for seven languages: Japanese, Korean, Russian, Turkish, English, French, and Italian. Of the languages in this sample, Japanese, Korean, and Turkish are predominantly SOV and strongly head-final, while the other languages are predominantly SVO and moderately to strongly head-initial (Liu, 2010; Futrell et al., 2015b). These languages were selected in order to explore whether the pattern found in Futrell et al. (2015a) and Futrell et al. (2020), where head-final languages tended to have longer dependencies than head-initial languages, holds in spoken dependencies, as well.

It has been argued (e.g. Hawkins, 2014; Futrell et al., 2020) that shorter dependencies make comprehending and producing language more efficient.

[1] https://support.google.com/youtube/topic/9257536?hl=en&ref_topic=9257610

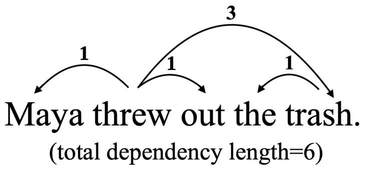

Figure 1: Dependency relations and lengths for the sentence "Maya threw out the trash." The length of a dependency is defined as the number of words intervening between a head and its dependent, plus the dependent itself. The total dependency length is the sum of these individual lengths, here 6. Figure adapted from Futrell et al. (2015a).

I thus hypothesized that writing would show less minimization than informal speech, which, given its more transitory nature (Auer, 2009), may create a pressure for shorter dependencies.

1.1 Data and pre-processing

Spoken corpora for each language were constructed as outlined in Appendix A. UD corpora were selected based on the corpus used to train Stanza's (Qi et al., 2018) default model for each language (Table 1). Due to the spoken corpora's skew toward short sentences, only sentences up to length 15, or approximately two standard deviations from the mean of the spoken dataset as a whole (μ=5.33, SD=4.46), were selected for analysis.

Language	Corpus	Sentences
Japanese	GSD	2715
Korean	GSD	3352
Russian	SynTagRus	29269
Turkish	IMST	3050
English	EWT	7620
French	GSD	4915
Italian	ISDT	6613

Table 1: Universal Dependencies 2.6 corpora selected for comparison to YouDePP corpora.

Following Futrell et al. (2015a), observed dependency lengths were calculated as the sum of the lengths of all dependency arcs in a sentence (Figure 1). A random baseline and optimized baseline were additionally calculated for comparison to the observed data. 10 random linearizations were generated for each observed sentence by recursively randomizing the order of each head and its dependents (Futrell et al., 2015a). Optimized linearizations were generated via the following algorithm (Gildea and Temperley, 2007):

1. Order the dependents of each head by weight, where weight is defined as the total number of children contained under the dependent.
2. Place the dependents on alternating sides of the head in order of weight, such that the heaviest dependents are the farthest from the head.
3. Finally, the link between the head and its parent node is treated as a special child that is placed opposite the head's longest real child.

1.2 Results

Qualitatively, both the observed data and baselines were highly similar across YouDePP and UD corpora (Appendix A). Additionally, a similar overall pattern to Futrell et al. (2015a) and Futrell et al. (2020) was observed, with Japanese as a notable exception: Turkish and Korean were both much closer to the random baseline than the other languages in this sample. In absolute terms, the dependencies in the YouDePP corpora were slightly longer than those in the UD corpora; this may be due to the greater variability in dependency lengths at each sentence length in the spoken corpus.

To determine whether the relative difference in growth rates between the random baseline and the observed dependency lengths differed significantly between spoken and written corpora, a linear mixed-effects model was fit for each language[2][3]. Following the methods of Futrell et al. (2015a) and Futrell et al. (2020), models were fit with total dependency length as the dependent variable, and squared sentence length[4], baseline (random or observed) and, additionally, corpus as predictors. Random intercepts by sentence were also included.

Significance testing was performed via model comparison. A significant three-way interaction

[2] The growth rates of dependency lengths in the manually-transcribed captions were significantly slower than those of the automatic captions for all three languages at $p < 0.02$ ($\beta_{it} = -0.002$, $\beta_{fr} = -0.008$, $\beta_{en} = -0.002$). The models reported here were run with automatic and manual data combined. Results of the mixed models did not change when only manual captions were included.

[3] The relative difference between random and observed baselines between corpora, rather than simply the difference between observed baselines between corpora, is used in order to control for factors that may affect all baselines of a given corpus together.

[4] Futrell et al. (2015a) found that using squared sentence length as a predictor provided a better fit to the data than linear sentence length.

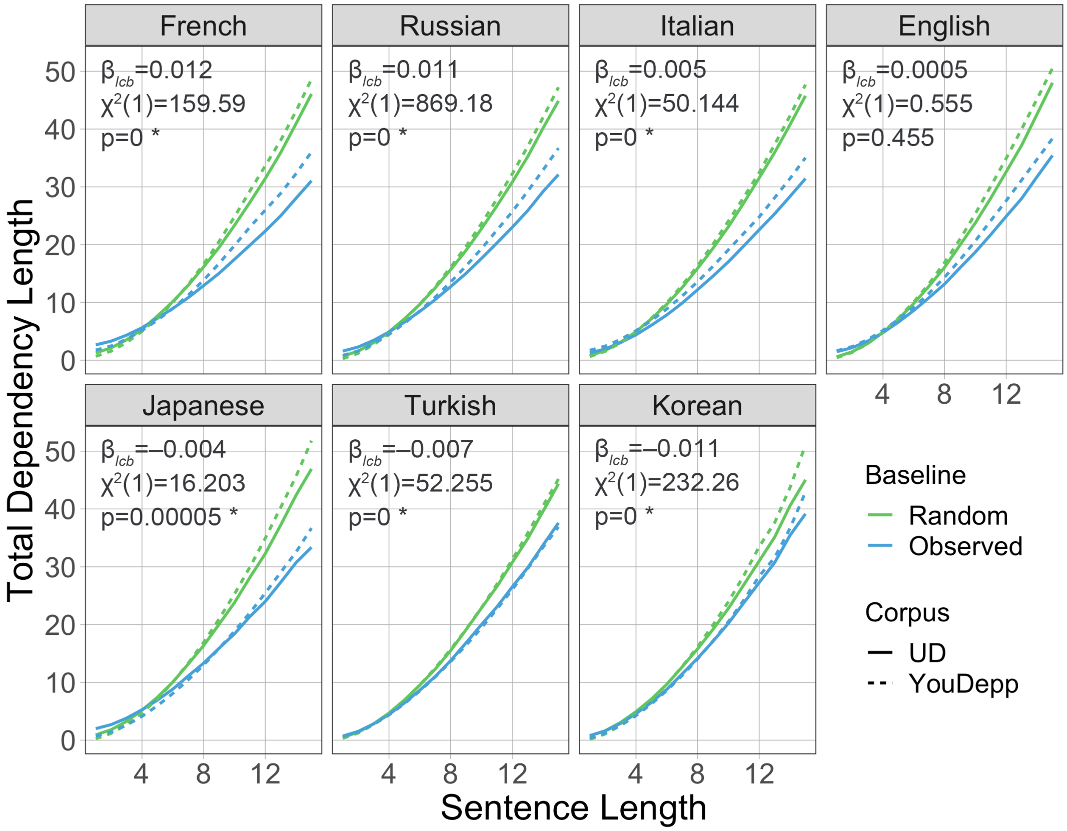

Figure 2: Model fits for random and optimal baselines in Universal Dependencies and YouDePP corpora. Sentence length is plotted along the x-axis and predicted mean total dependency length for all sentences of a given length is plotted on the y-axis. Colored lines represent random and observed baselines. Dashed lines indicate predictions for YouDePP, and solid lines represent predictions for UD. The β term is the coefficient of the interaction between squared sentence length (l), baseline (b), and corpus (c). This coefficient indicates how quickly the observed baseline grows relative to the random baseline in each YouDePP corpus as compared to the corresponding UD corpus. Languages are ordered by the size of this coefficient, i.e., the difference in degree of minimization between observed dependencies and random dependencies across the two corpora, from longer spoken dependencies than written dependencies (French, Russian, Italian; English no difference) to shorter spoken dependencies than written dependencies (Japanese, Korean, Turkish).

between sentence length, baseline, and corpus indicates that the growth rate of the observed baseline, as compared to the relative baseline, differs significantly between the two corpora for a given language. Significance values for each language and predicted dependency lengths for each combination of language, baseline, and corpus are shown in Figure 2. Observed dependencies in Italian, French, and Russian grew at a significantly faster rate relative to random dependencies in the YouDePP (spoken) corpora than in Universal Dependencies (written) corpora, while English dependencies grew at nearly the same rate relative to the random baseline across the two corpora. In contrast, dependencies in Japanese, Korean, and Turkish grew at a slower rate relative to their random baselines in the spoken corpora than in the written corpora.

2 Discussion

Counter to my predictions, dependency lengths were not universally shorter in speech than in writing. Furthermore, the SOV, head-final languages in this sample did not show a markedly different pattern from that in Futrell et al. (2015a), with the exception of Japanese. The overall patterning of dependency lengths was in fact highly similar across all corpora. That being said, it is possible that the similarity seen here is in fact an artifact of the way these languages are tagged. Further work should aim to confirm whether these patterns hold independently.

However, Japanese, Turkish, and Korean did pattern together in an interesting way: they were the only languages in the sample to show a significantly higher degree of minimization with respect to sentence length in speech than in writing. One possible interpretation of this result is that, because the dependency lengths in SVO and/or more head-initial languages are already closer to "optimal" in writing than those of SOV and/or strongly head-final languages, deviations from written patterns in speech are more likely to result in longer dependencies than in SOV/head-final languages. In contrast, Japanese, Korean, and Turkish allow for more flexible word order and more extensive argument drop in speech than in writing. Particularly in the case of Korean and Turkish, the observed dependencies of which are much closer to the random baselines than other languages, deviations from written norms may be more likely to lower dependencies than increase them.

That being said, there was much more variability in the YouTube data than in the UD data. It is likely that the extremely long dependencies seen at higher sentences lengths are the result of strings of, e.g., repeated words or other nonsensical text that the parser cannot handle well. Removal of these problematic strings could potentially bring dependency lengths down in the spoken corpora for all languages; this is another area in which additional work is needed.

3 Conclusion

Comparison of spoken and written dependency lengths using corpora gathered via YouDePP and corpora from Universal Dependencies 2.6 found that, overall, dependency lengths in speech patterned similarly to those in writing. Furthermore, counter to expectations, spoken dependency lengths were not consistently shorter than those in writing. Instead, languages varied in whether they minimized more or less in speech than in writing, and this variation patterned with headedness, such that more head-initial, predominantly-SVO languages minimized dependencies less in speech, while more head-final, predominantly-SOV languages minimized them more.

It is possible that the SOV languages in this sample, particularly Korean and Turkish, have more room to minimize in speech compared to writing. Notably, these languages also make greater use of flexible word order and argument drop in speech than in writing, which may have more of an impact on dependency lengths in SOV contexts. In order to better understand how features such as argument drop and word order flexibility contribute to dependency length minimization, future work aims to extend this method to languages that vary systematically with respect to canonical order, head/dependent marking, and other potentially relevant features.

References

Peter Auer. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, 31(1):1–13.

Douglas Biber. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2):133–163.

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

Steven Coats. 2020. Articulation Rate in American English in a Corpus of YouTube Videos. *Language and Speech*, 63(4):799–831.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96:371–412.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015a. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015b. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics*, pages 91–100.

Daniel Gildea and David Temperley. 2007. Optimizing Grammars for Minimum Dependency Length. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, page 8.

John A. Hawkins. 2014. *Cross-Linguistic Variation and Efficiency*, first edition edition. Oxford Linguistics. Oxford University Press.

Stig Johansson and Knut Hofland. 1994. Towards an English–Norwegian parallel corpus. In Udo Fries, Gunnel Tottie, and Peter Schneider, editors, *Creating and Using English Language Corpora*, page 25–37. Editions Rodopi.

Natalia Levshina. 2017. Online film subtitles as a corpus: An n-gram approach. *Corpora*, 12(3):311–338.

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

Shigeko Nariyama. 2000. Referent identification for ellipted arguments in Japanese.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Paulo Quaglio. 2008. Television dialogue and natural conversation: Linguistic similarities and functional differences. In Annelie Ädel and Randi Reppen, editors, *Corpora and Discourse: The Challenges of Different Settings*, pages 189–210. John Benjamins.

Mieko Ueno and Maria Polinsky. 2009. Does headedness affect processing? A new look at the VO–OV contrast. *Journal of Linguistics*, 45(3):675–710.

Tatiana Zdorenko. 2010. Subject omission in Russian: A study of the Russian National Corpus. In Stefan Th. Gries, Stefanie Wulff, and Mark Davies, editors, *Corpus-Linguistic Applications*, pages 119–133. Brill — Rodopi.

Daniel Zeman, Joakim Nivre, and Mitchell Abrams et al. 2020. Universal dependencies 2.6. LINDAT/ CLARIAH-CZ, Faculty of Mathematics and Physics, Charles University.

A Corpus construction

The top YouTube channels in Japan, Korea, Russia, Turkey, the United States, France, and Italy were manually determined on the basis of total subscribers via the ranking websites https://www.noxinfluencer.com/, https://socialblade.com/, and https://vidooly.com. Only channels categorized as "Entertainment," "Comedy," or "People and Blogs" were selected for processing. These categories were chosen because they tended to contain content that was highly informal and conversation-heavy. Video URLs were scraped from each channel's "Videos" page via Selenium[5]. Original-language captions were then obtained using pytube[6]. Initially, only manually-transcribed captions were selected.

There was considerable variation in the number of manually-captioned videos available for a given language. For languages that had a high proportion of manually-captioned videos, the top five channels (by number of subscribers) featuring a significant amount of dialogue were selected for further processing. For languages that had few manually-captioned videos, channels were scraped in order of rank (again by number of subscribers) until at least 100 manually-captioned videos were found.

The results of this initial collection process were somewhat surprising: popular channels in Japan tended to contain a relatively high proportion of manually-captioned videos, channels from Korea, Russia, and Turkey contained a moderate number, and channels from the United States, France, and Italy had very few manually-captioned videos. It is possible that speakers of languages for which auto-captioning is more accurate are less likely to caption videos; it is also possible that popular channels in major world languages, particularly English, are less likely to make use of YouTube's community captions feature due to abuse.

To supplement the sparse data from English, Italian, and French, auto-generated captions were also collected. Auto-generated captions were taken from the top five channels for which manually-transcribed captions were available, with a cap of

[5] https://www.selenium.dev/

[6] https://github.com/nficano/pytube

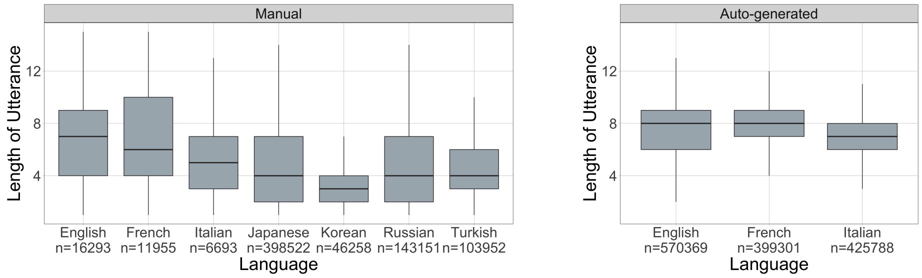

Figure 3: *Left.* Distribution of utterance lengths in manually-transcribed captions by language for utterances up to length 15. *Right.* Distribution of utterance lengths in auto-generated captions by language for utterances up to length 15.

	Manual captions		Automatic captions	
Language	Videos	Lines	Videos	Lines
Japanese	2767	403145	–	–
Korean	800	48208	–	–
Russian	726	157091	–	–
Turkish	505	156558	–	–
English	241	16293	1631	570369
French	122	11955	1280	399301
Italian	102	6693	7708	425788

Table 2: Total number of videos and lines (utterances) with manual and automatic captions scraped by language.

up to 500 captions per channel. The total number of manual and automatic caption files collected for each language, as well as the number of utterances contained in these files, are presented in Table 2. In general, sentences from both automatic and user-contributed sources were very short; the distribution of sentence lengths in the range from 0–15 dependencies are given in Figure 3.

Downloaded captions were pre-processed to remove symbols and emojis, as well as to correct unusual punctuation (for example, the use of tildes in place of periods) or lack of punctuation[7]. Additional processing was done to remove speaker attributions, parentheticals, and other extraneous text, such as sound effects. After pre-processing, the captions were parsed using Stanza (Qi et al., 2018). The default packages were used for all languages. The total number of utterances per language after pre-procesing and parsing are given in Table 2.

There are some important caveats to keep in mind when using this method to automatically process and parse YouTube captions. First, captioners do not all follow the same conventions. In particular, some write one complete sentence per line, while others break up sentences across lines. Of those who break up sentences across lines, very few use commas or periods, making it difficult to determine automatically whether a given line is part of a previous sentence or the start of a new sentence.

Here, periods were automatically added to any lines that did not already end with a period, comma, or other form of punctuation. In other words, lines that were ambiguous between being a complete utterance or a part of a longer utterance were treated as complete utterances. Furthermore, naturalistic spontaneous speech is often fairly fragmented and contains a high number of one-word utterances (Biber, 1992). As a result, the dataset is heavily skewed toward very short utterances (Figure 3). The differences in sentence lengths across languages are not necessarily a reflection of actual differences between the languages; they could also be a function of content and/or captioners' styles.

In addition, although using the same parser across languages has the benefit of consistent tagging, the parser does not perform equally well with each language. This is of particular issue for languages with very different spoken and written forms, such as Japanese. Spoken Japanese is more flexible than written Japanese, which is rigidly verb-final, and particles, which encode part-of-speech information, are frequently omitted. Parsers trained on written Japanese thus tend to mis-parse non-verb-final orders and unmarked nouns.[8]

[7] Emojis, creative punctuation, and other symbols are sometimes used in YouTube captions to represent prosody or speakers' emotions.

[8] In these cases, the overall dependency structure is still largely correct—particle-less nouns, for example, are correctly treated as dependents of the verb, even if their part of speech is tagged incorrectly. Thus, these mis-parses may not be an issue for coarse-grained analyses of dependencies, but may need to be corrected for finer-grained analyses to be feasible.

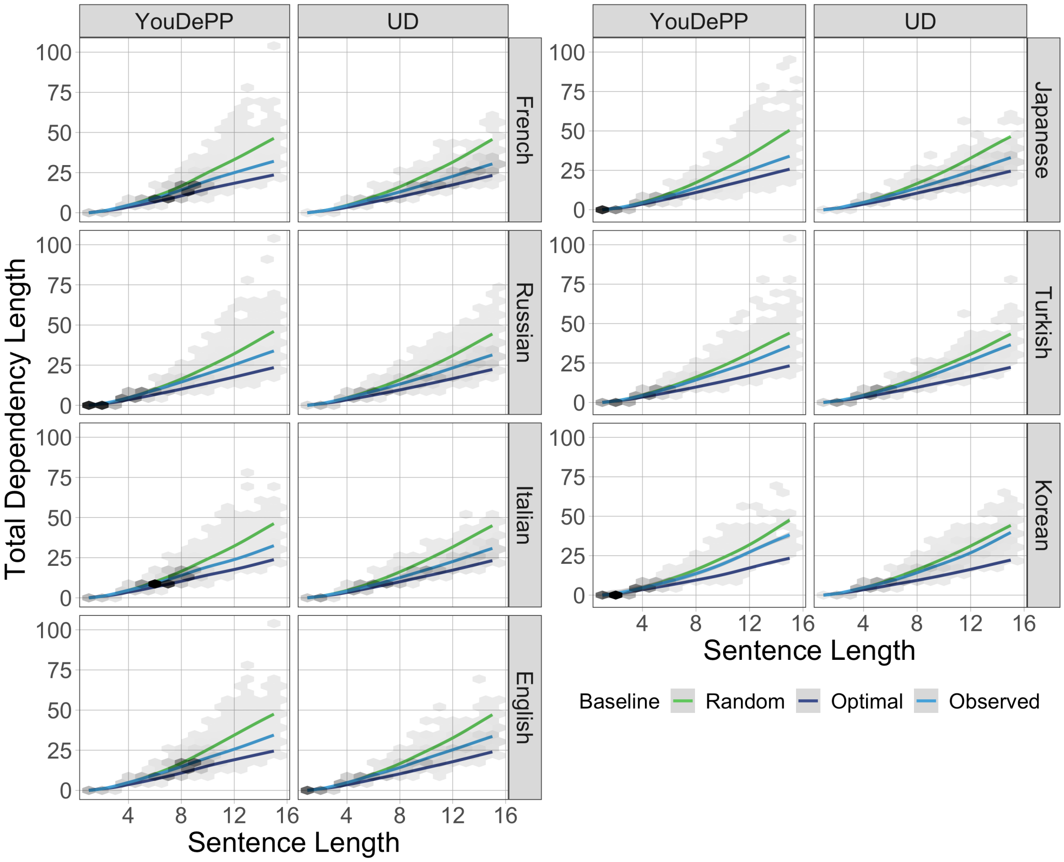

Figure 4: Total dependency length as a function of sentence length in the YouDePP corpora and Universal Dependencies 2.6. Sentence length is plotted along the x-axis and mean total dependency length for all sentences of a given length is plotted on the y-axis. Colored lines, fit with a generalized additive model for visualization, represent random, optimal, and observed baselines. Hexagons represent the density of observed sentences at each combination of sentence length and dependency length. Languages are ordered by the difference in degree of minimization between observed dependencies and random dependencies across the two corpora, from longer spoken dependencies than written dependencies (French, Russian, Italian; English no difference) to shorter spoken dependencies than written dependencies (Japanese, Korean, Turkish).