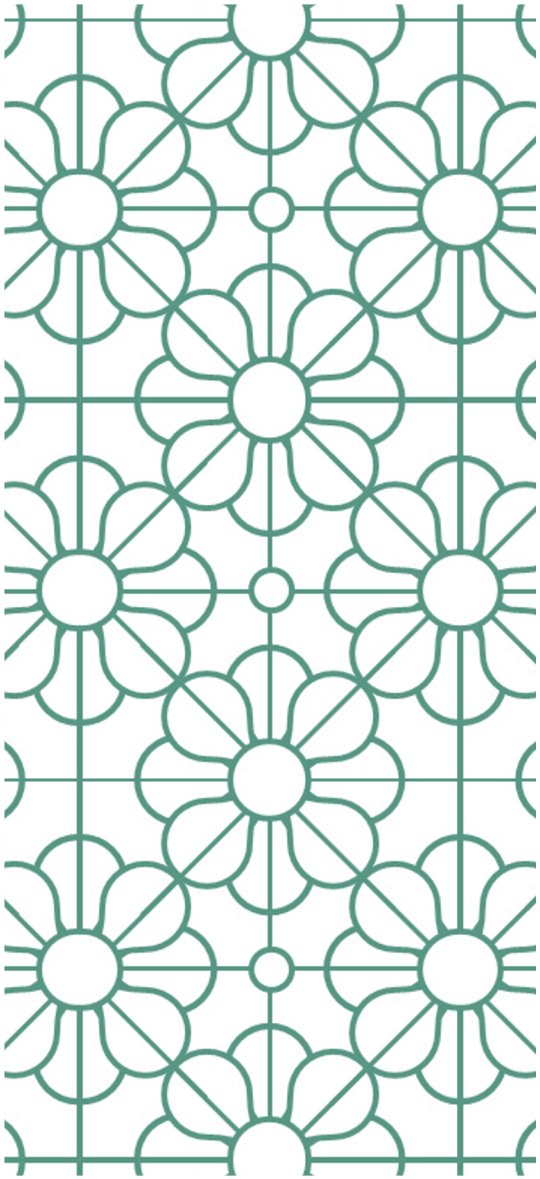


THE EFFECT OF REGISTER ON DEPENDENCY LENGTH IN TWO FLEXIBLE LANGUAGES

Alex Kramer
University of Michigan
SLE 2020: New Perspectives on Word Order
8/27/2020

OVERVIEW

Big questions

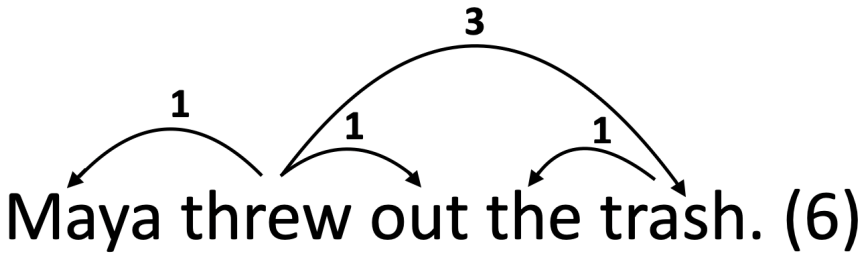
- How and why does the degree of dependency length minimization differ between languages? Between registers?
- What new sources of (informal) speech data can we use to help answer this question?

Approach: Dependency-parsed YouTube captions (YouDePP)

DEPENDENCY LENGTH AND DEPENDENCY LENGTH MINIMIZATION

Dependency length

- The distance (# of words) between a dependent and its head



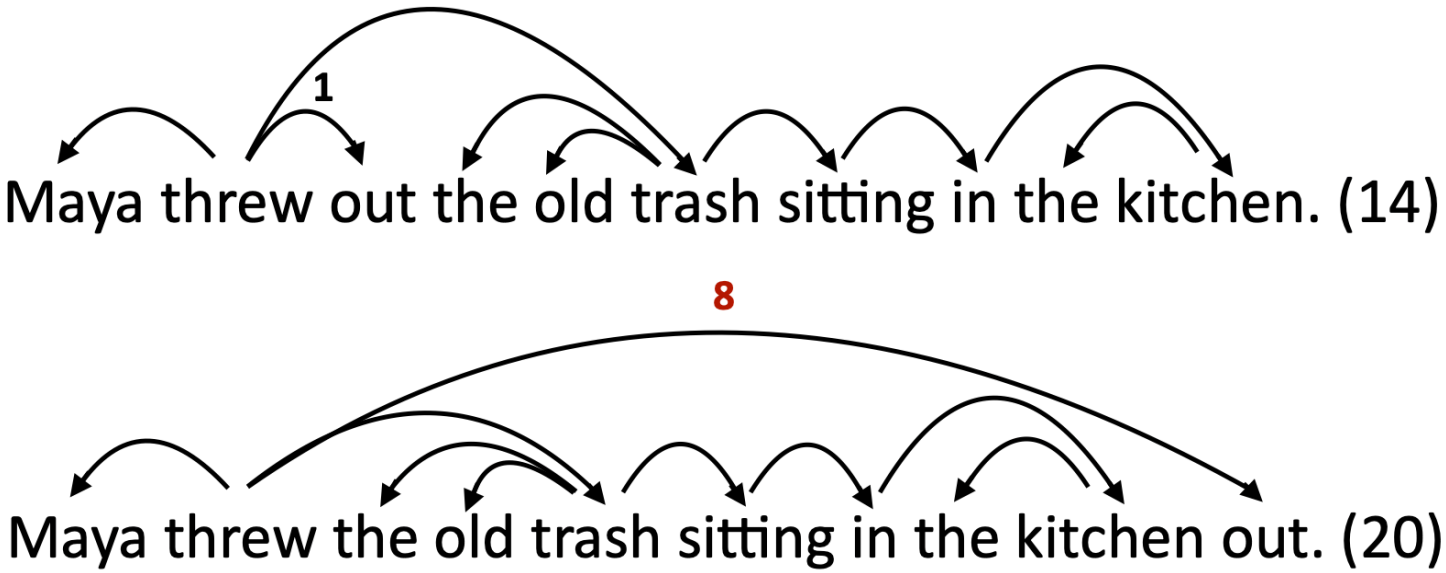
(Adapted from Futrell et al. 2015a)

Dependency length minimization hypothesis

- "The evolution of languages is driven by the constraint that grammars should allow dependents to be realized as closely as possible to their heads" (Yu et al. 2019)

Minimization strategies available to languages can differ!

MINIMIZATION STRATEGIES: WORD ORDER

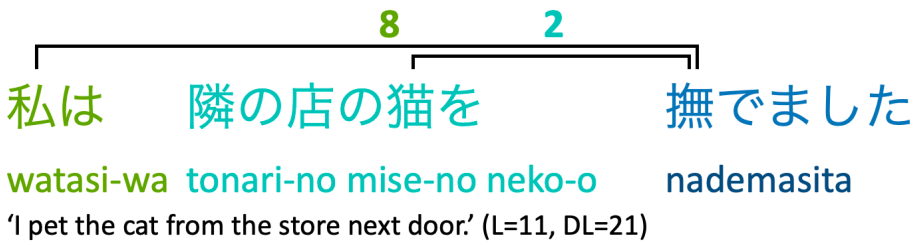


(Adapted from Futrell et al. 2015a)

MINIMIZATION STRATEGIES: ARGUMENT DROP

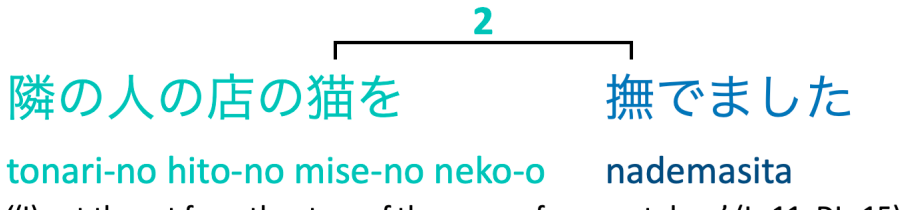
- Has been discussed as a feature that can reduce absolute length (Ueno & Polinsky 2009)
- Can also make a difference for dependency length
- As good as changing word order!

(1)



'I pet the cat from the store next door.' (L=11, DL=21)

(2)

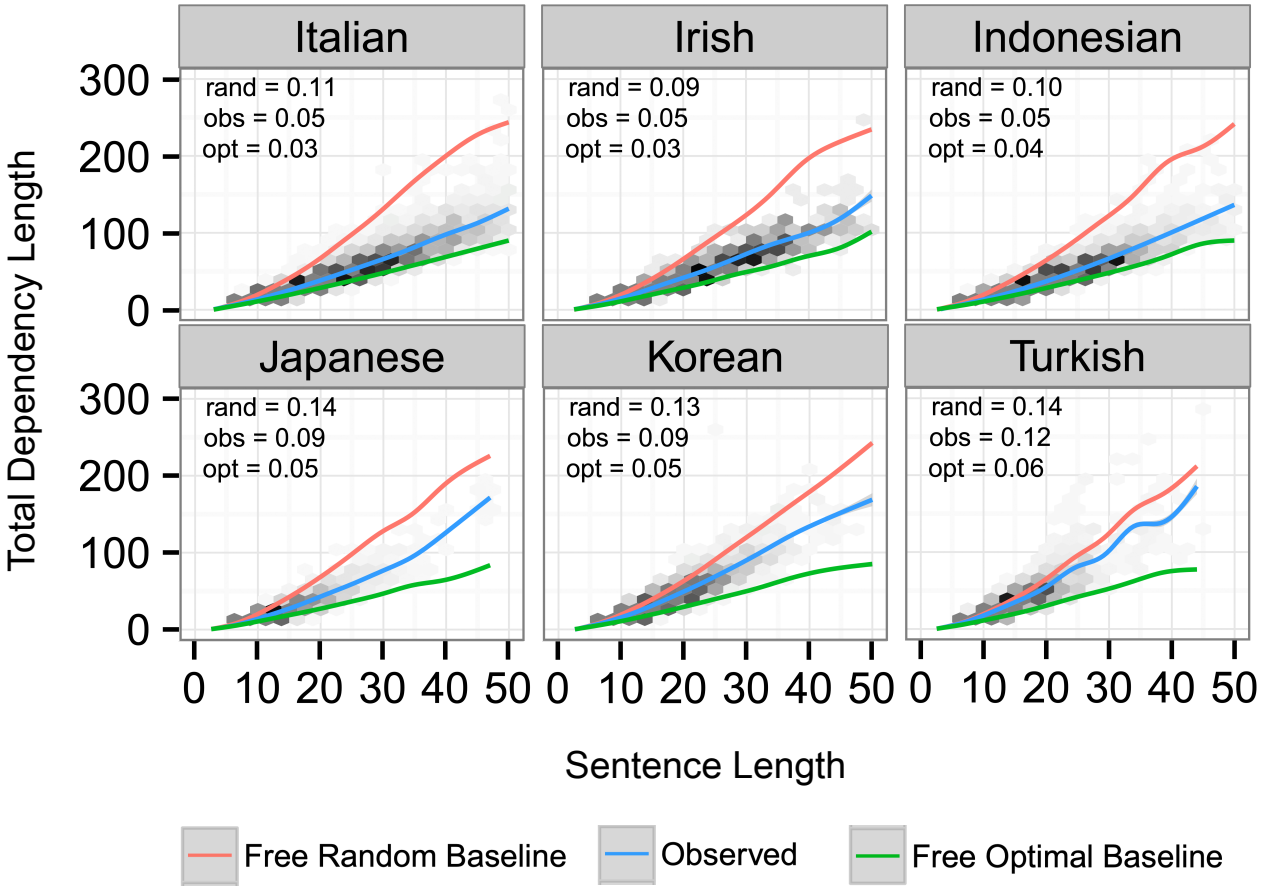


'(I) pet the cat from the store of the person from next door.' (L=11, DL=15)

(3)

'I pet the cat from the store next door.' (L=11, DL=16)

VARIATION IN DLM: PREVIOUS WORK



- Languages such as Japanese, Korean, Turkish have been found to have longer average dependency lengths than, e.g., Italian, Indonesian, Irish (Futrell 2015a)
- Why?
  - Headedness?
  - Constituent order?
  - Flexibility?

(Figure from Futrell et al. 2015a)

DEPENDENCY CORPORA: SOME ISSUES

- Many large corpora historically use written language & formal registers
- But, there are systematic differences between registers of a language (Biber 1993)
- By extension, cross-linguistic corpus comparisons aren't necessarily comparing the same kinds of language use

QUESTIONS REVISITED

- Flexible word order and argument drop can both reduce dependency lengths
- Both strategies are available in Japanese, especially casual spoken Japanese

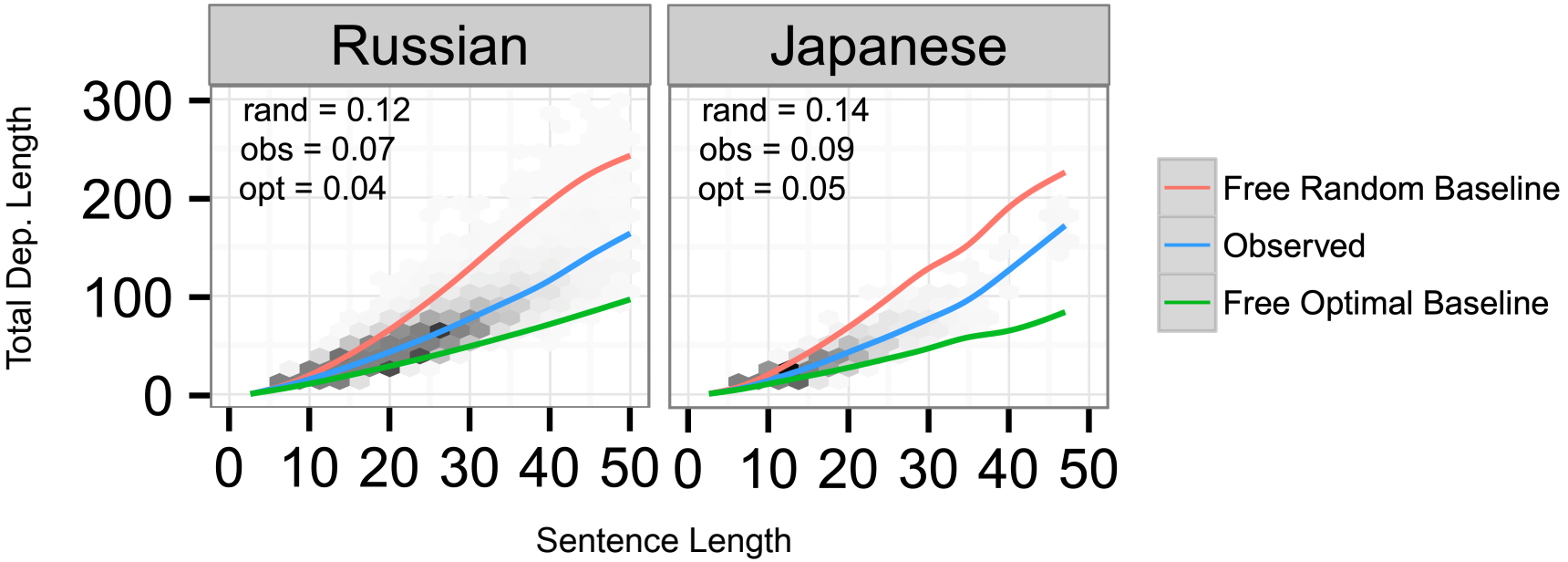
---

- Will very casual spoken Japanese look different from previous results?
  - If so, how do these features contribute to DLM in casual speech?
- How does casual spoken Japanese compare to casual spoken Russian, a flexible SVO language with more limited argument drop?

JAPANESE VS. RUSSIAN: FLEXIBILITY & ARGUMENT DROP

Japanese	Russian
▪ Flexible SOV	▪ Flexible SVO
▪ Writing rigidly verb-final Casual speech optionally flexible	▪ Writing and speech flexible Non-SVO orders better in some contexts
▪ Frequent argument drop	▪ Limited argument drop
▪ Subject drop (Nariyama 2000) 20% Novels 37% News 74% Conversation	▪ Subject drop (writing) (Zdorenko 2010, Seo 2001) 0~22% Fiction 2% Blogs 4% News
▪ Written arg. drop (Ueno & Polinsky 2009) 12% Politics 22% Mystery 25% Magazines	▪ Subject drop (speech) (Zdorenko 2010, Grenoble 2001) 3% Lectures 6% Interviews 24~32% Conversation, stories

JAPANESE VS. RUSSIAN: PREVIOUS DLM RESULTS



Russian: SynTagRus (Droganova et al. 2018)

- Written texts from various genres: news, fiction, blogs, etc.

Japanese: Tüba J/S (Heinrichs et al. 2000)

- Spoken dialogues from 3 formal situations: Appointments, travel, computer maintenance

(Figure from Futrell et al. 2015a)

PREDICTIONS

- Both Japanese and Russian will show more minimization in YouTube data, i.e., informal speech, than in previous work
- Japanese will show a greater difference in minimization between informal and formal registers than Russian due to the difference in availability of argument drop & flexibility

METHODS: CORPUS OF YOUTUBE CAPTIONS

Two types of captions:

Auto-generated (speech-to-text)

- Doesn't work well on non-English languages

Community-provided

- By and large of good quality

YouTube might discontinue the community captions feature this year!
Sign the petition!

METHODS: CORPUS GENERATION

1. Identify top YouTube channels in Japan and Russia that contain speech
2. Scrape captions from videos using `PyTube` module
3. Automatically process captions to remove things that trip up parser
   - In progress: Manual processing of ~10% of data for comparison
4. Parse with `stanza` [`StanfordNLP`] (Qi et al. 2018)

METHODS: SOME LIMITATIONS

- Parser doesn't perform well with casual speech, especially Japanese
  - Lack of particles
  - Flexible order
- That said, it usually gets the dependency structure right *enough*, since what matters here is distance

METHODS: DEPENDENCY CALCULATION

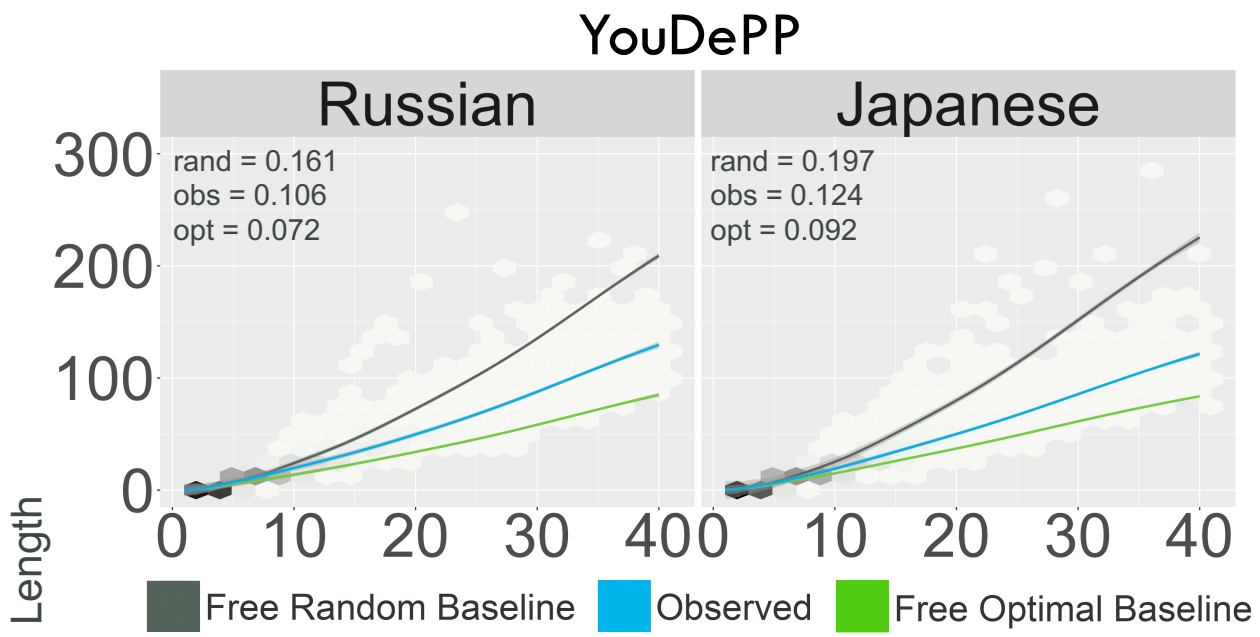
For each sentence:

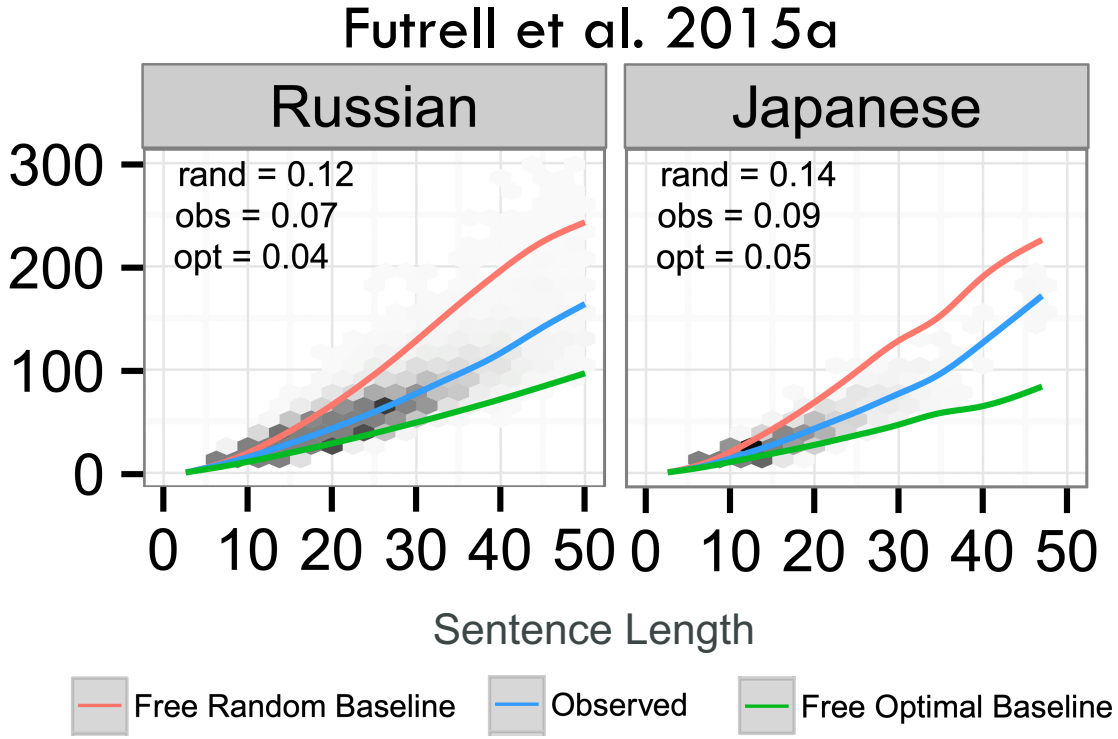
1. Calculate total dependency length of the sentence (sans punctuation)
2. Generate 10 random linearizations of the sentence and calculate total dependency length for each
3. Generate optimal arrangement of sentence and calculate total dependency length (Gildea & Temperly 2007)

RESULTS: DEPENDENCY LENGTH GROWTH RATE

Minimization ratio:

$$\frac{\text{area}_{\text{rand}} - \text{area}_{\text{obs}}}{\text{area}_{\text{rand}} - \text{area}_{\text{opt}}}$$





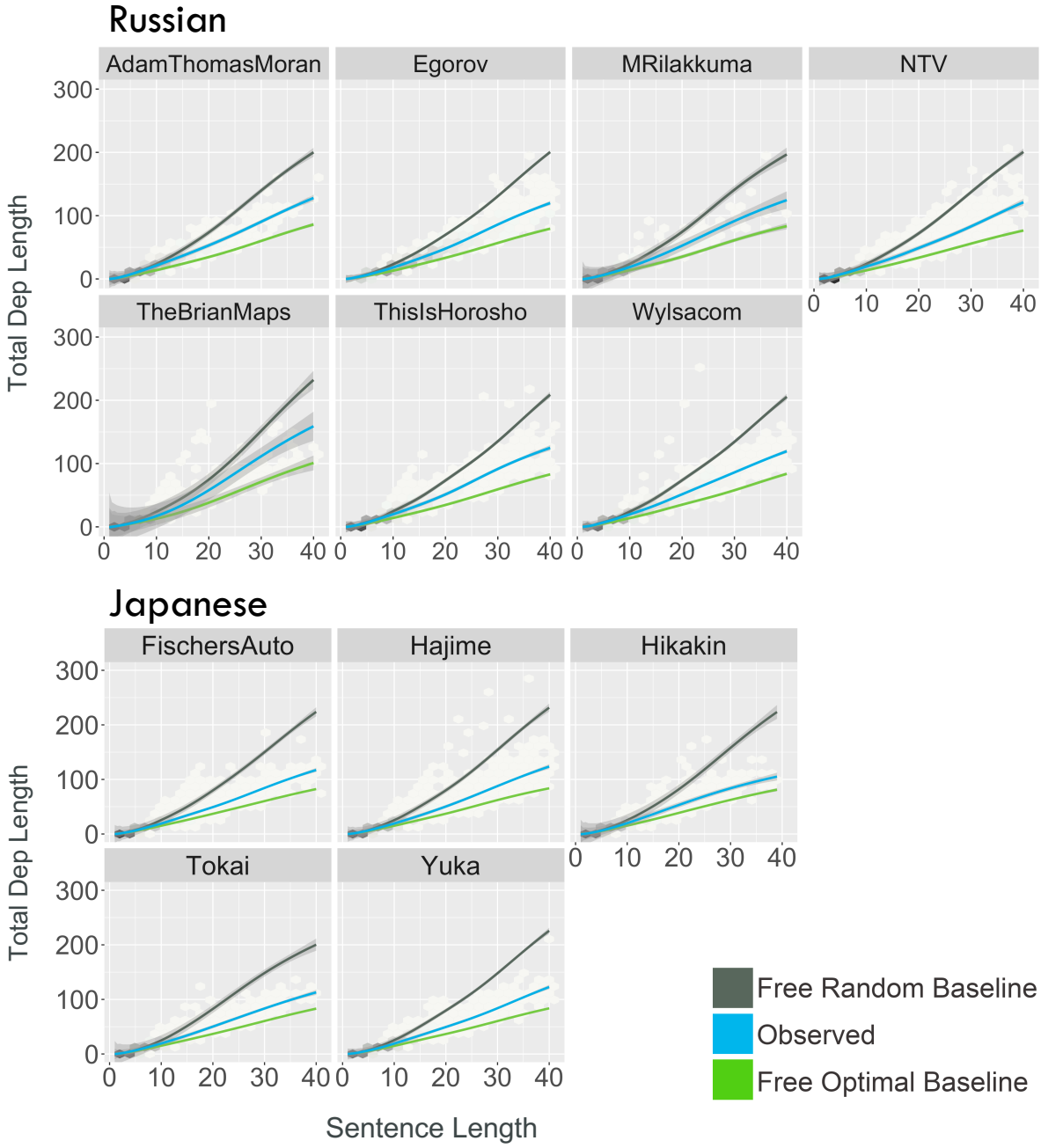
RESULTS: MINIMIZATION

Language	Corpus/Study	Minimization Ratio
Japanese	YouDePP	0.689 *(0.695)*
	Futrell et al. 2015	0.556
Russian	YouDePP	0.606 *(0.618)*
	Futrell et al. 2015	0.618

- Casual spoken Japanese minimizes much more than formal spoken data
- Casual spoken Japanese minimizes more than casual spoken Russian
- Russian is similar across the two studies

RESULTS: BY-CHANNEL COMPARISON

- It's possible that channels of different genres would show different trends
- No clear pattern in Russian data
- More monologue-heavy Japanese channels seem similar?
- Need to compare more channels and genres!



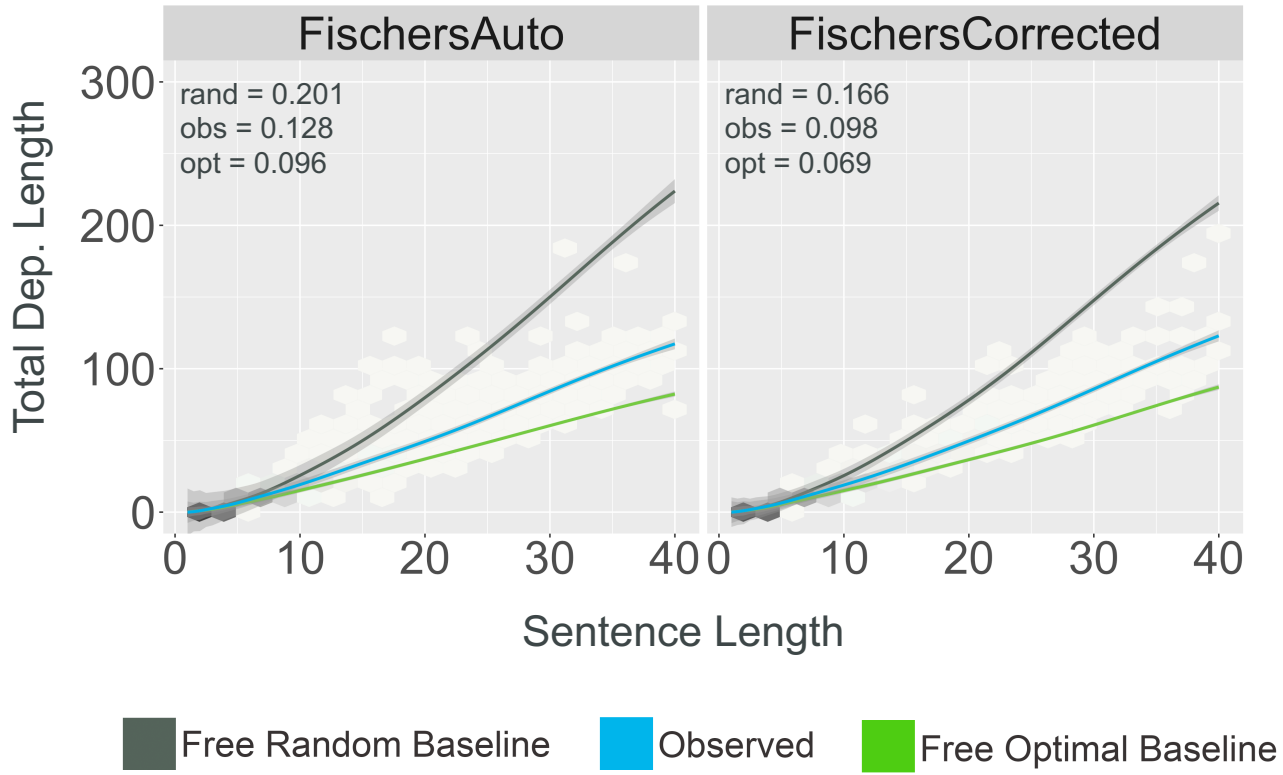
METHODS: HAND-CORRECTIONS

In conjunction with watching video:

1. Determine sentence boundaries
2. Remove nonsense lines (e.g. sound effects, laughter)
3. "Correct" slang forms/spellings
4. Note use of non-canonical orders, code for argument drop

RESULTS: HAND CORRECTIONS

- Growth of all baselines significantly slower
- Overall pattern unchanged
  - **Auto ratio:** 0.691
  - **Corrected ratio:** 0.696
- More stable estimates at higher sentence lengths



RESULTS: ARGUMENT DROP VS. ORDER IN JAPANESE

Only about 5% (363/7500) of hand-checked utterances use a non-canonical order

In contrast, arg. drop is extremely common

- 76% of clauses drop at least one argument
- Quick comparison—frequency of 1st person singular pronouns in Japanese vs. Russian data:
  - RU (я): ~39 times/video; 16.2% of sentences
  - JA (*[w]ata[ku]si, boku, ore*): ~5 times/vid; 3.36% of sents

CONCLUSIONS

- Argument drop may be driving dependency length minimization more than flexible word order in casual spoken Japanese
- These strategies allow casual spoken Japanese to minimize dependencies more than casual spoken Russian

MOVING FORWARD

- Manual correction and annotation
- **More languages & registers**
  - Variety of written and spoken sources

- Looking at (informal) speech through new mediums is an important complement to existing written data
- Sources like YouTube can serve as powerful tools for uncovering typological patterns that are difficult to detect when we only look at formal registers and written modalities

THANK YOU!

Sign the petition!

REFERENCES

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133–163. https://doi.org/10.1080/01638539209544806

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

Dyachenko, P. V., Iomdin, L. L., Lazursky, A. V., Mityushin, L. G., Podlesskaya, O. Y., Sizov, V. G., Frolova, T. I., & Tsinman, L. L. (2015). A Deeply Annotated Corpus Of Russian Texts (SynTagRus): Contemporary State Of Affairs [Sovremennoe Sostojanie Gluboko Annotirovannogo Korpusa Tekstov Russkogo Jazyka (SinTagRus)]. *Proceedings of the VV Vinogradov Russian Language Institute [Trudy Instituta Russkogo Yazyka Imeni VV Vinogradova]*, *6*, 272–299.

Futrell, R., Mahowald, K., & Gibson, E. (2015a). 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

Futrell, R., Mahowald, K., & Gibson, E. (2015b). Quantifying Word Order Freedom in Dependency Corpora. *Proceedings of the Third International Conference on Dependency Linguistics*, 91–100.

Grenoble, L. A. (2001). "Conceptual reference points, pronouns, and conversational structure in Russian." *Glossos*, 1(1).

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency* (First edition). Oxford University Press.

Hinrichs, E. W., Bartels, J., Kawata, Y., Kordoni, V., & Telljoann, H. (2000). The Verbmobil treebanks. In E. G. Schukat-Talamazzini & W. Zühlke (Eds.), *Sprachkommunikation* (pp. 107–112). VDE-Verlag.

REFERENCES

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

Levshina, N. (2019). *Token-based typology and word order entropy: A study based on Universal Dependencies*.

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*, 160–170. https://doi.org/10.18653/v1/K18-2016

Seo, S. (2001). "The frequency of null subject in Russian, Polish, Czech, Bulgarian, and Serbo-Croatian: an analysis according to morphosyntactic environments." Ph.D. dissertation, Indiana University at Bloomington.

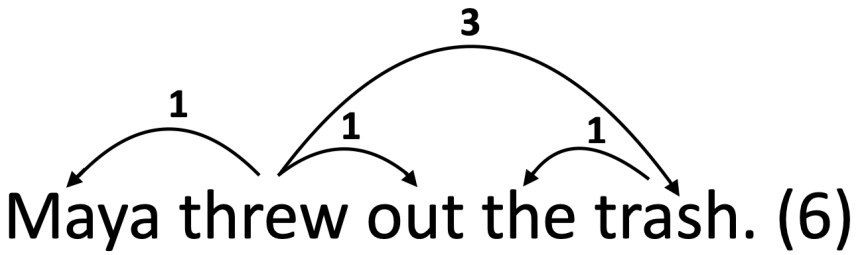
Ueno, M., & Polinsky, M. (2009). Does headedness affect processing? A new look at the VO–OV contrast. *Journal of Linguistics*, *45*(3), 675–710. https://doi.org/10.1017/S0022226709990065

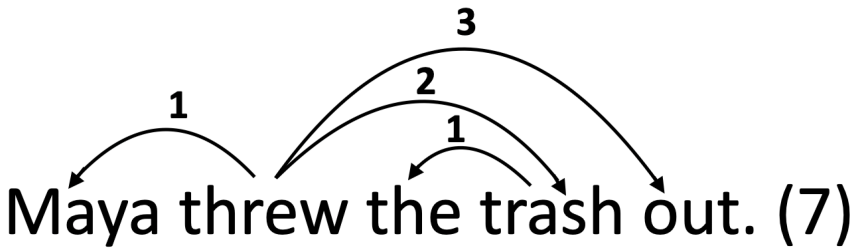
Yu, X., Falenska, A., & Kuhn, J. (2019). Dependency Length Minimization vs. Word Order Constraints: An Empirical Study On *55* Treebanks. *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, 89–97. https://doi.org/10.18653/v1/W19-7911

Zdorenko, T. (2010). Subject omission in Russian: A study of the Russian National Corpus. In S. Th. Gries, S. Wulff, & M. Davies (Eds.), *Corpus-linguistic applications* (pp. 119–133). Brill | Rodopi. https://doi.org/10.1163/9789042028012_009

MINIMIZATION STRATEGIES: WORD ORDER

Case 1: Different orders have similar lengths





(Adapted from Futrell et al. 2015a)

VARIATION IN DLM: WHY?

Possible sources of variation:

- Headedness?
  - JA, TR and KO are SOV and strongly head-final (Liu 2010, Futrell 2015b, Levshina 2019)
  - IT and ID (SVO) and Irish (VSO) are moderately–strongly head-initial (Liu 2010, Futrell 2015a,b)
  - More freedom in dependency lengths due to, e.g., identifiability or lack of overlap between forms (Futrell 2015a, Hawkins 2014, Levshina 2019)?
- Some constructions "naturally" longer (SOV will be longer than equivalent SVO)?—but then we might expect trade-offs elsewhere
- Flexibility?

RESULTS: DEP LENGTH GROWTH RATE