Revisiting dependency length and intervener complexity minimisation on a parallel corpus in 35 languages

Andrew Dyer

Universität des Saarlandes, Germany

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

Introduction

Dependency Length Minimisation (DLM)

- Word order choice to minimise distance between a dependant and its head.
- Widely seen as a universal, observed in all languages in a sample of 37 by Futrell et al. (2015).

However,

- Appears less strong or even absent in verb-final languages.
- Has been proposed to be a by-product of **intervener complexity measure** (ICM) reduction (Yadav et al., 2022).
  - Minimisation of syntactic heads between a dependant and its head.

(a) Intervener complexity = 1,

Dependency length = 4

(b) Intervener complexity = 3,

Dependency length = 4

Figure: Illustration of dependency length in contrast with intervener complexity. Image borrowed from Yadav et al. (2022).

Introduction: Universal Dependencies

Previous studies mostly use **Universal Dependencies**. This has several advantages for comparative study:

- ✓ High quality manual annotation
- ✓ Wide language coverage

But also some disadvantages:

- ✘ Domain and content differences between languages
- ✘ Different datasizes between languages

In other words, are differences in the language corpus contents confounding results?

- A more general question in quantitative typology.

Introduction: Parallel Corpora

An alternative is to use **parallel corpora** such as Parallel Universal Dependencies (PUD), or parsed Bible corpora.

- ✓ Parallel sentences ensure comparability
- ✓ Comparable datasizes
- ✘ Often narrow language coverage or small corpus size (PUD)
- ✘ May contain highly specific lects (Bible)

Corpus: CIEP+

Our approach:
Corpus of Indo-European Prose Plus (CIEP+) (Talamo and Verkerk, 2022).

- ✓ Coverage of 37 languages (30 IE, 7 non-IE) so far
- ✓ Oriented towards more natural-sounding translations, modern language
- ✘ Unknown extent of Translationese.
- ✘ Limits us to a set of mostly LOL (Dahl, 2015) languages.

Method

We follow the experimental setup of Futrell et al. (2015).

- Compare dependency lengths in sentences in 35 languages to baseline **permutations** of these sentences.
- Measure the **rate of dependency length increase** for each baseline.
- Same method for **intervener complexity measure**.

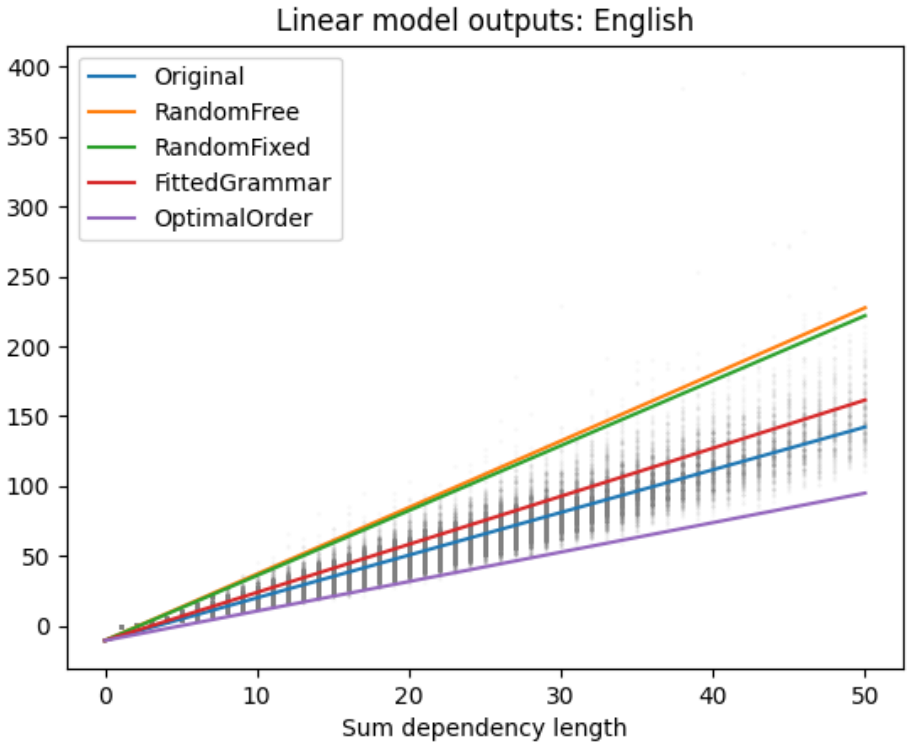
Permutation modes

- **RandomFree**: Random projective permutation.
- **RandomFixed**: Permutation according to randomly generated grammars.
- **FittedGrammar**: Permutation according to grammar obtained by corpus counts.
- **OptimalOrder**: Permutation that optimises for dependency length; based on Gildea and Temperley (2007).

Permutation mode	**Aleatory**	**Fixed order**	**Futrell et al.**
RandomFree	✓		✓
RandomFixed	✓	✓	✓
FittedGrammar		✓	
OptimalOrder			✓

Data analysis

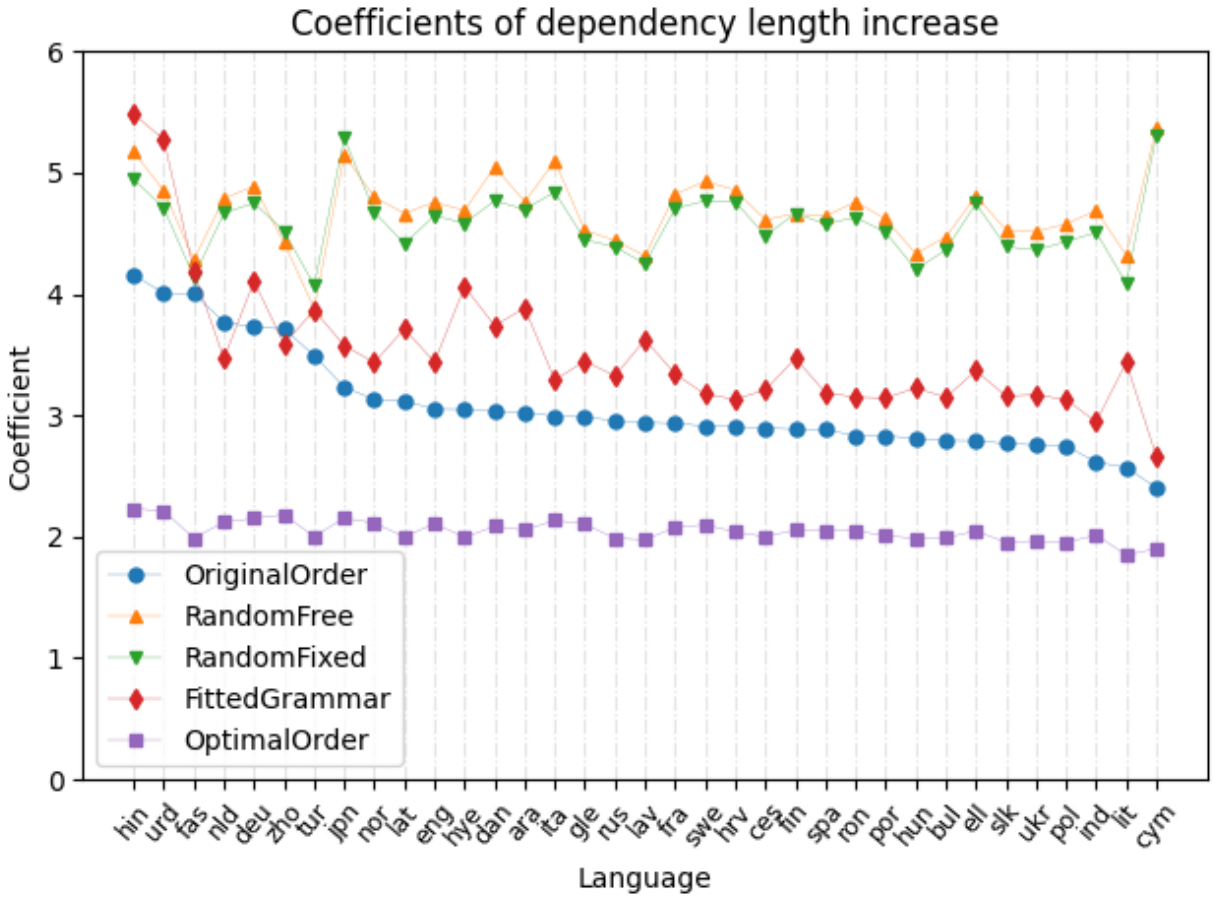
Linear Mixed-Effects Regression (LMER) to find the coefficient for dependency length increase by sentence length, for each permutation mode.



LMER formula

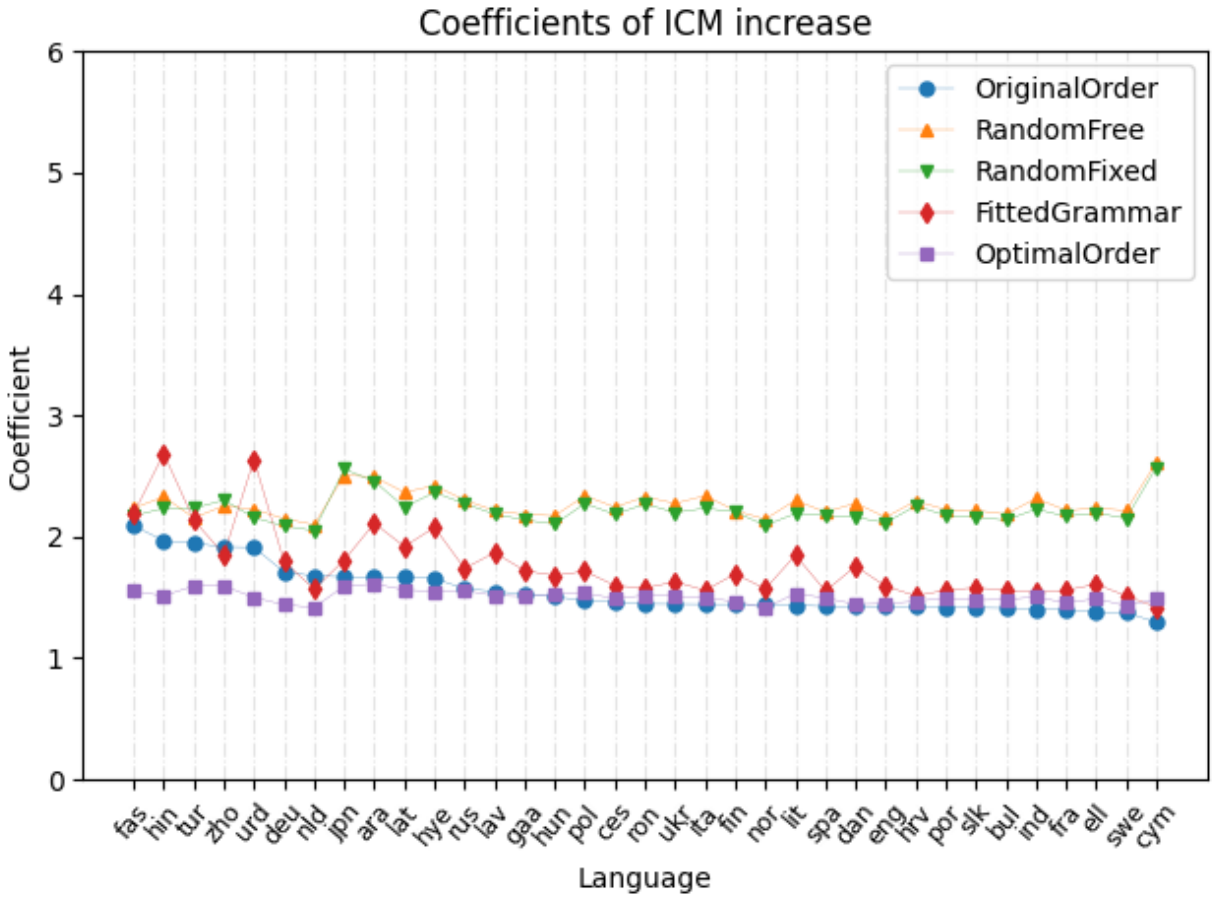
$$\texttt{DependencyLength} \sim \texttt{SentenceLength} * \texttt{PermutationMode} + (1|\texttt{ID})$$

Results: Dependency Length



- **OriginalOrder** is well below random baselines in all languages, though less so in verb-final languages.
- **FittedGrammar** is also well below the random baselines, though not as low as OriginalOrder.
- **OptimalOrder** is consistent between languages.

Results: Intervener Complexity Measure



- ICM is also minimised compared to the random baselines.
- A similar asymmetry is visible with verb-final languages.
- **OriginalOrder** is close to or even below **OptimalOrder**.

Discussion

Our results with a parallel corpus broadly mirror the findings of Futrell et al. (2015) with UD. How to interpret this?

- DLM shows through the noise of domain variation.
- Because we look at languages *as a whole*, variation evens out across sentence type.

Parallel corpora may be more important for studies targeted towards specific construction types.

Conclusion

- DLM as a universal is upheld in our study with a parallel corpus.
- The asymmetry in verb-final languages is also evident in this study.
- DLM is achieved in part by both canonical orderings and word order flexibility.
- ICM appears to be close to fully optimised, but shows the same verb-final asymmetry.

We hope to use CIEP+ in further studies to great effect.

Dahl, O. (2015). How weird are wals languages? In *Diversity Linguistics: Retrospect and Prospect*.

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

Gildea, D. and Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191.

Talamo, L. and Verkerk, A. (2022). A new methodology for an old problem: A corpus-based typology of adnominal word order in european languages. *Italian Journal of Linguistics*, 34:171–226.

Yadav, H., Mittal, S., and Husain, S. (2022). A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, 6:147–168.

Key studies

There are two particular studies of interest here:

- Futrell et al. (2015)
  - Compares observed sentence dependency lengths to random permutations of the same sentences.
  - Finds clear DLM effect in all 37 languages in the sample.
  - **But** the effect is weaker in verb-final and V2 languages.
- Yadav et al. (2022)
  - Introduces the concept of Intervener Complexity Measure (ICM): the number of heads between a dependant and its head.
  - Does not compare ICM between languages directly; evaluates the explanatory value of the metric using all languages as a random effect.