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

Andrew Dyer
Saarland University
Language Science & Technology
andrew.dyer@uni-saarland.de

Abstract

In this replication study of previous research into dependency length minimisation (DLM), we pilot a new parallel multilingual parsed corpus to examine whether previous findings are upheld when controlling for variation in domain and sentence content between languages. We follow the approach of previous research in comparing the dependency lengths of observed sentences in a multilingual corpus to a variety of baselines: permutations of the sentences, either random or according to some fixed schema. We go on to compare DLM with intervener complexity measure (ICM), an alternative measure of syntactic complexity. Our findings uphold both dependency length and intervener complexity minimisation in all languages under investigation. We also find a markedly lesser extent of dependency length minimisation in verb-final languages, and the same for intervener complexity measure. We conclude that dependency length and intervener complexity minimisation as universals are upheld when controlling for domain and content variation, but that further research is needed into the asymmetry between verb-final and other languages in this regard.

1 Introduction

Efficiency in language production and processing is widely held as a universal, underpinning various aspects of human language evolution and use. (Levshina and Moran, 2021). Within syntax, an expression of this is found in the theory of Dependency Locality (Gibson, 1998): the principle that syntactically related information should appear in close proximity in a sentence, so as to minimise the memory load required to parse it. Its observable effect is dependency length minimisation (DLM): the ordering of a sentence such that the sum distance of dependencies in sentences is minimised (Gibson). This effect has been well studied cross-lingually (Gildea and Temperley, 2010; Liu, 2008), and is widely held to be a universal of syntax, with the study of Futrell et al. (2015) finding evidence of the effect in all languages in a sample of 37 languages in Universal Dependencies (Nivre et al., 2016), among other such cross-lingual studies.[1]

There remain, however, inconsistencies and asymmetries in how and where DLM is applied, within languages and within sentence structures. For example, a common finding is that DLM is less pronounced or even absent in head-final languages when controlling for various factors such as sentence type, and when looking only at lexical tokens. Jing et al. (2022) find a negative association between head-finality and dependency length when controlling for harmony and considering only lexical dependencies, an effect that they find to be robust against multiple random baselines. Liu (2021) also finds mixed evidence for the correlation between dependency length and ordering choices for pre-verbal arguments in head-final languages; whereas argument ordering choice is more clearly associated with dependency length in languages with post-verbal arguments. These findings point to a more nuanced picture of DLM, where the effect is asymmetric in terms of word order, more clearly pronounced in head-initial languages (Yadav et al., 2020).

Another question is the extent to which DLM exists as an independent effect, as opposed to being a function of other constraints. Yadav et al. (2022) propose the alternative measure of sentence complexity, Intervener Complexity Measure (ICM), which measures not the number of tokens between dependants and their heads, but the number of syntactic heads between them, suggesting optimisation for ICM underlies the observed DLM effect.

As a first step in investigating these questions, our replication study revisits the work of Futrell and Gibson (2015) to broadly replicate this study on a new corpus, with some additions in light of

[1] https://universaldependencies.org/

subsequent research. We seek to reevaluate the following questions:

1. Does the observation of DLM in all languages hold when languages contain loosely parallel data?
2. To what extent is DLM achieved by word order variation, as opposed to canonical word order constraints?
3. Do we see the same asymmetry between verb-final and non-final languages as in previous works?
4. How does DLM compare to ICM minimization across languages?

Our study pilots a new corpus: the Corpus of Indo-European Prose Plus, or CIEP+ (Talamo and Verkerk, 2022). CIEP+ is a parallel corpus of translated works of modern prose in several languages, syntactically annotated under the Universal Dependencies[2] framework. The translated texts are drawn from the most widely translated works of prose in the world. While the corpus originated as a means of comparative study of Indo-European languages, and these languages make up the majority of its data, it also contains translations in some non-Indo-European languages.

The use of parallel corpora is beneficial in making language data more comparable between languages, controlling for domain differences and the natural variation of communicative intent in sentences (Dahl, 2007). However, most currently available parallel corpora suffer either from limited size and language coverage (e.g. Parallel Universal Dependencies), or from being drawn from highly specific lects that do not reflect common language use (e.g. parallel Bible corpora, UN Declaration of Human Rights).

A related problem in parallel corpora is the phenomenon of *Translationese* (Gellerstam, 1986): the effect whereby translated texts are identifiable by certain characteristics that are atypical in the target language, caused by language-specific or universal effects of the translation process (Koppel and Ordan, 2011).

Fictional and non-fictional prose are not immune to the effects of Translationese (Puurtinen, 2003; Popescu, 2011). Nevertheless, since the goal of translated prose is entertainment rather than exactitude, we expect that translators will use stylistic translations that may be closer to the conventions of the target language, thus mitigating this concern.[3]

The books that we use are large, containing thousands of sentences. And, though we do not escape the bias of translations mostly being available in a small set of languages, we nevertheless manage a decent coverage of 35 languages, with at least 20,000 sentences in each.

Our use of this corpus addresses a potentially confounding issue in Futrell et al. (2015) and other corpus-based studies: the variation in domain coverage across the UD corpora. With a parallel corpus, we once again put these findings to the test.

2 Background

Word ordering with respect to phrase heaviness has long been a topic of interest in constituency-based syntax (Arnold et al., 2000), and has been adapted to a dependency grammar framework as dependency length (Gildea and Temperley, 2010).

Since the inception of Universal Dependencies (Nivre et al., 2016) and other consistently annotated multilingual corpora, more multilingual studies of DLM have been carried out. Futrell and Gibson (2015) compare the sum dependency lengths of observed sentences in 37 languages in Universal Dependencies to random baselines of sentences permuted to random orders. They find that in all 37 languages, dependency length as a function of sentence length shows a consistently slower increase than would be expected in random word order baselines, whether free or fixed.

Yu et al. (2019) extend this study to probe the impacts of canonical word order constraints versus variability on DLM. Building on the setup of Futrell and Gibson (2015), they use randomly permuted baselines with *same valency* (i.e. all heads in the permuted sentence must have the same number of dependants on each side) and *same side* (i.e. dependants must be on the same side of their head) constraints, and find that each baseline shows a reduction in dependency length, and that atypical orderings in a language usually contribute to this.

In several studies, Liu (2020, 2021, 2022) probes the DLM effect with regard to ordering flexibility and pre- and postverbal argument domains. Among her findings is that while dependency length minimisation is well-correlated with phrase ordering

[2] https://universaldependencies.org

[3] We are unaware of any quantitative evaluation of the prevalence of Translationese in prose compared to other genres of translation, such as legal, technical and political translations. Such research would be very valuable.

choices in postverbal languages (e.g. English, Bulgarian, Dutch), this effect is much weaker or nonexistent in preverbal languages (e.g. Japanese, Persian), suggesting that the relevance of DLM depends greatly on word ordering constraints, among other pressures.

Intervener Complexity Measure is introduced by Yadav et al. (2022). They operationalise the complexity of intervening information in long dependencies as Intervener Complexity Measure, which counts the number of syntactic heads between a dependant and its head. By comparing random permutations of trees alternately matched for dependency length or intervener complexity, they find that random linear arrangements matched for dependency length tend to have very close ICM to the original sentence, but that the inverse effect is not as strong. Though Yadav et al. (2022) perform their experiments using several languages in Surface Universal Dependencies (Gerdes et al., 2018), accounting for language as a random effect, we are unaware of any multilingual study so far that has directly measured the extent of intervener complexity minimisation per language.

Most prior large cross-lingual studies of dependency length minimisation have used Universal Dependencies or Surface Universal Dependencies corpora, or other dependency corpora pre-dating UD (Liu, 2008), without control for domain and sentence variation. However, there are some that have used parallel corpora. For example, Jiang and Liu (2015) compare effects of sentence length and dependency direction in a parallel English-Chinese corpus; and Ferrer-i Cancho (2017) use the Parallel Universal Dependencies (PUD) corpora. We are unaware of any previous work with parallel corpora of the same size as CIEP+.

3 Method

In our investigation, we broadly replicate the experimental setup of Futrell and Gibson (2015).

The dependency length of a token in a sentence is defined as the number of tokens between it and its head in the linear surface order, including itself (i.e. a minimum of 1). The dependency length of a sentence is then the sum of dependency lengths for each token, excluding the root.

We compare the dependency lengths of observed sentences to a set of random baselines: reorderings of the sentences in the corpora with the same underlying tree structure but a different linear surface order of tokens. These baselines are:

1. **RandomFree** Random projective permutations of the sentence retaining the same structure.
2. **RandomFixed** Permutations according to a randomly generated grammar.
3. **FittedGrammar** Permutations of each sentence to strictly follow an approximation of the language's canonical word order.
4. **OptimalOrder** Permutation of each sentence to optimise for minimum dependency length.

Of these, FittedGrammar is introduced by our study, while the others are also used by Futrell and Gibson (2015). We briefly describe and motivate each permutation method in Section 3.1.

After creating permutations of each sentence in each book in each language, we use a linear mixed-effects model to estimate the rate at which dependency lengths increase as a function of sentence length. The response variable of the model is sentence dependency length, while the fixed variables are the interaction between sentence length (in number of tokens) and permutation mode: the baseline that produced the sentence (including the unaltered original sentence).

We use sentence ID as a random effect in the model. Sentence ID is shared across all permutations of a sentence, and including it accounts for the effect of the variance in sentence structure. This random effect is simplified compared to Futrell and Gibson's, which groups permutations by sentence ID. We found that doing this caused singular fits in the model.

Performing this separately for each language in the corpus, we use the coefficient of the model fit as the measure of a language's rate of dependency length increase. The higher the coefficient, the greater the dependency lengths we can expect to see as sentence length increases. The model gives us a separate fit for each of the baselines, and so we are able to compare the true rate of increase to what we could expect to see in each of the baseline conditions. If the true rate of increase is not lower from the random baseline, for example, then we do not see DLM in the language.

We use the same approach to measure intervener complexity minimisation. The intervener complexity of a token is defined by Yadav et al. (2022) as the number of syntactic heads that come between it and its own head; including the token's head itself, meaning that for each token the minimum intervener complexity is one. The Intervener Complexity Measure of a sentence is then the sum of tokenwise intervener complexities in the sentence. Fig. 1 shows an example of Intervener Complexity Measure for a sentence in contrast to dependency length.

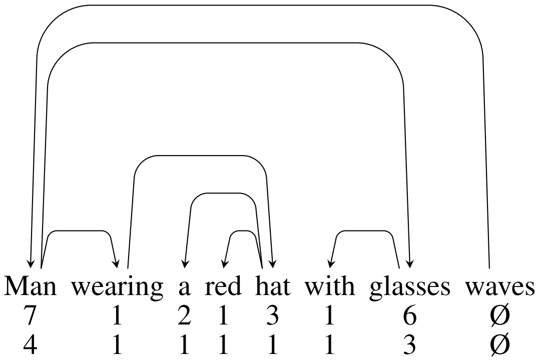

Figure 1: A demonstration of the difference between dependency length and intervener complexity. The top layer of numbers is dependency length; the bottom layer is intervener complexity.
For example. there are seven tokens between *waves* and its syntactic head *Man*, but only three heads between them (*wearing*, *hat* and *glasses*).
The ICM of this sentence is 12, compared to 21 for dependency length.

3.1 Permutation baselines

RandomFree

In the RandomFree baseline, we recursively permute each subtree within a sentence tree such that the children of any head may appear in any order before or after the head. The same underlying tree structure is retained, but the linear surface order is random with the sole constraint that the resulting tree is projective. We perform this procedure 10 times for each sentence in the corpus.[4]

If DLM holds, then the observed dependency lengths should be consistently below what we would expect to see in random linear arrangements of the same sentence.

RandomFixed

We use the term *grammar* throughout this paper to refer to a lookup table for a determinate position of each dependency relation with respect to its head.

For each dependency relation, we assign a lookup value in the range [-1,1]. For each recursive subtree in the corpus, the dependants are rearranged according to the lookup value of their dependency relation. Dependencies whose label has a negative lookup value go to the left of their head; those with a positive value go to the right. The higher the absolute value, the further the sentence is from the head in the new sentence permutation.

As in the RandomFixed baseline, we produce 10 random grammars in total, and permute each sentence according to each of these grammars.

This baseline is a more conservative variant of the random free baseline, taking into account that all languages have at least some degree of fixedness in their word order, the regularity of which is hypothesised to reduce dependency length on average.

FittedGrammar

The fitted grammar for each language is a count-based estimation of the majority position for each dependency relation. For each dependency relation, we assign two parameters: $sign$ - an integer -1 or 1, depending on whether the dependency relation most often appears on the left (-1) or the right (1) of its head; and $distance'$ - a float of the mean log distance of the dependency relation from its head (relative to other dependants) when on the side indicated by $sign$. The final parameter $position$ is then the product of $sign \times distance'$: a positive or negative real number. As in the random fixed baseline, all dependants are then ordered according to this lookup value. Fig. 2 shows an example of how such a grammar would assign the order of dependants.

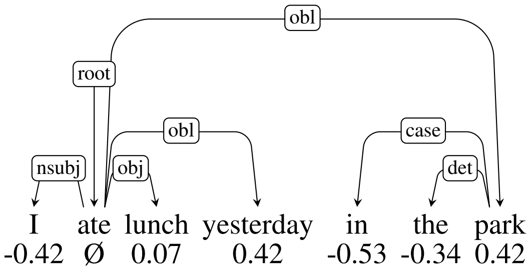
obl
root
obl
case
nsubj
obj
det
I ate lunch yesterday in the park
-0.42 Ø 0.07 0.42 -0.53 -0.34 0.42

Figure 2: An example of how a grammar might assign the positions of dependants. Below each word is the position lookup value for its dependency relation. For example, `nsubj` has a $position$ value of -0.42. When two dependants have the same lookup value (as in *yesterday* and *park* here) the ordering of the two is arbitrary. The lookup values in this example are taken from the fitted grammar for English.

The fitted grammar is used as a rough measure

[4] Futrell and Gibson's setup calls for 100 random permutations. We find that this number quickly becomes intractable for storage and processing with our larger corpus size.

of the extent to which DLM is achieved through language users' choice of sentence orderings as opposed to the canonical word order constraints of the language. We find that the lookup values obtained by this method generally match with canonical word order classifications.

For example, Table 1 shows some lookup values for nsubj, obj and obl relations in four languages. In each of these languages, the relative lookup values correspond with the orderings of subject, object, and verb (SOV) (Dryer, 2013) and oblique, object and verb (XOV) (Dryer and Gensler, 2013) in WALS. Though we cannot fully model the canonical word order rules of a language with only the basic relations of UD, we can at least provide an approximation that is comparable between languages.

	nsubj	**obj**	**obl**	**WALS**	
				Ch. 81	**Ch. 84**
eng	-0.42	0.07	0.42	SVO	VOX
jpn	-0.60	-0.14	-0.52	SOV	XOV
ara	0.18	0.59	0.55	VSO	VOX
zho	-0.77	0.39	-0.51	SVO	XVO

Table 1: The *position* values for nsubj, obj and obl in four languages. For example, in Japanese the obl relation has a lower value than obj, meaning that it will be placed before it; and both have a negative value, so they will both be placed to before their head. Assuming that the head is a verb, this follows the canonical XOV word order in Japanese.

OptimalOrder

Our algorithm for finding the optimal linear order that minimises dependency length is based on that of Gildea and Temperley. For each recursive subtree, we sort dependants by their *weight*: the number of words in their recursive subtree. Dependants are then placed inside-out on alternating sides of their head. Whether the alternation starts from left or right depends on the direction of the head: left-branching heads will start left-to-right; right-branching heads, right-to-left. This order will be reversed if the number of dependants is even, such that the heaviest dependant will branch in the same direction as its head. Fig. 3 shows an example of the output of this algorithm.

The optimal ordering gives an idea of the upper bound of DLM that we could expect under complete word order freedom with DLM as the only objective. In the case of languages with a high dependency length rate, the comparison with the optimal baseline tells us to what extent this can be explained by the inherent complexity of the sentence structure.

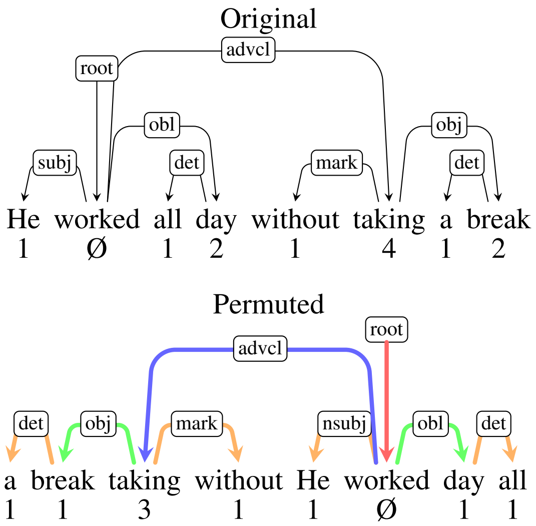

Figure 3: An example of how the OptimalOrder algorithm permutes a sentence. The colour of the edges indicates the order in which they are attached to their head: orange first; green second; blue third. Dependency lengths are shown on the bottom row of text. The permuted sentence has lower dependency lengths than the original due to the flattening effect of the algoithm.

3.2 Data

We use the CIEP+ corpus for our analysis (Talamo and Verkerk, 2022).[5] CIEP+ is a parallel corpus of translated works of modern prose in several languages, comprised of a set of some of the world's most widely translated works. The source languages of the texts varies between English, French, Portuguese, Spanish, German, and Dutch. The corpus is parsed predictively using the Stanza NLP pipeline (Qi et al., 2020), which has pretrained models in UD format with Labeled Attachment Score of at least 70% for all languages under consideration. The languages that we use, their families, and their canonical word order are shown in Table 2. These languages are not subset to those used by Futrell et al. (2015) and thus cannot be directly compared, but are the languages for which we have data in CIEP+.

We remove all punctuation tokens from the corpus, as these carry no semantic information and cause artificially long dependency lengths. In or-

[5] https://www.uni-saarland.de/fileadmin/upload/lehrstuhl/verkerk/CIEP_outline.pdf

Family	Language (code)	Basic order
IE Germanic	Danish (dan)	SVO
	Dutch (nld)	SVO*
	English (eng)	SVO
	German (deu)	SVO*
	Norwegian (nor)	SVO
	Swedish (swe)	SVO
IE Celtic	Irish (gle)	VSO
	Welsh (cym)	VSO
IE Romance	French (fra)	SVO
	Italian (ita)	SVO
	Latin (lat)	SVO
	Portuguese (por)	SVO
	Romanian (ron)	SVO
	Spanish (spa)	SVO
IE Baltic	Latvian (lav)	SVO
	Lithuanian (lit)	SVO
IE Slavic	Bulgarian (bul)	SVO
	Croatian (hrv)	SVO
	Czech (ces)	SVO
	Polish (pol)	SVO
	Russian (rus)	SVO
	Slovak (slk)	SVO
	Slovenian (slv)	SVO
	Ukrainian (ukr)	SVO
IE Indo-Iranian	Hindi (hin)	SOV
	Persian (fas)	SOV
	Urdu (urd)	SOV
IE Other	Armenian (hye)	SOV*
	Greek (ell)	SVO*
non-IE Finno-Ugric	Finnish (fin)	SVO
	Hungarian (hun)	SVO
non-IE Other	Arabic (ara)	VSO
	Chinese (zho)	SVO
	Indonesian (ind)	SVO
	Japanese (jpn)	SOV
	Turkish (tur)	SOV*

Table 2: Languages in CIEP+ tbat we use for our experiments. All languages have at least 20k sentences and are parsed using models with >70% LAS. Basic word order is according to WALS (Dryer, 2013). Asterisks * indicate that the language has more than one dominant word order.

der to reduce the number of parameters needed for the FittedGrammar and RandomFixed baselines, we simplify subtyped relations to their main type (e.g. `aux:pass` → `aux`). For ease of processing, we exclude non-standard tokens that are not part of the tree structure in the `conllu` format, such as enhanced dependencies and multiword tokens.[6] Finally, we exclude all sentences that, after these cleaning steps, have more than 50 tokens.

4 Results

4.1 Dependency length minimisation

We show the coefficients for the mixed-effects regression for each baseline in each language in Fig. 5. These coefficients represent the rate at which dependency length can be expected to increase as a function of sentence length for each baseline and language in the corpus. We also show an example of the regression fit in English in Fig. 4.

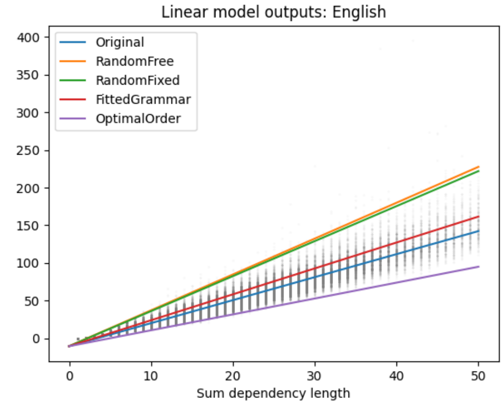

Figure 4: Dependency lengths as a function of sentence length in English. The coloured lines show the fit from the linear mixed-effects model for each baseline. Grey dots show the true (observed) dependency lengths.

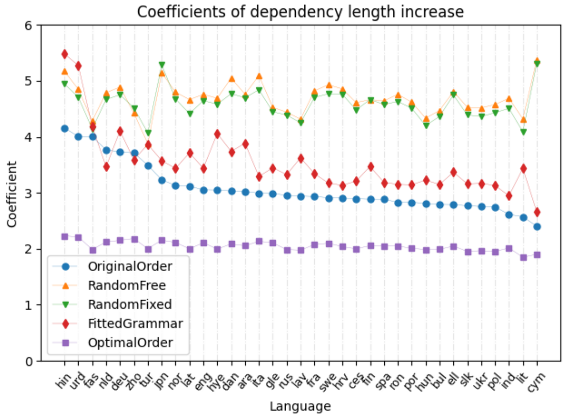

Figure 5: The coefficients of dependency length increase for all baselines in all languages. Languages are sorted in descending order by the coefficient of the OriginalOrder sentence.

Overall, we see clear evidence of DLM in all languages compared to both random baselines. We also find the same asymmetry as Futrell and Gibson (2015) and others whereby verb-final languages such as Hindi, Turkish and Japanese - and languages with frequent verb-final constructions such

[6] The reason for this is simply that such tokens are incompatible with our permutation algorithms. We leave examination of the impact of enhanced dependencies and multiword tokens on dependency lengths for future research.

as German, Dutch and Chinese - show faster rates of dependency length increase. We can see this as the rising tail on the left of the OriginalOrder coefficients in Fig. 5. The same tendency is not apparent in predominantly SVO languages with free word order and rich inflectional morphology, such as Baltic and Slavic languages.

Interestingly, we find the lowest rate of increase in Welsh, a VSO-preferring language (Williams, 1980), which we might expect to generate longer dependencies because of the increased distance from the predicate to its arguments. Irish, another Celtic language that prefers VSO word order, has a coefficient more in line with the SVO languages in the corpus. We should note that Welsh has one of the lower number of sentences in CIEP+, and the LAS of the Welsh parsing model in Stanza is low compared to other languages in our corpus, so we do not make any conclusions regarding this.

OptimalOrder is consistent across languages, showing that a consistent rate of increase is possible across all the languages sampled. This optimum would not be realistic in any of the languages as it would require no word order constraints, but it does show that where some languages show a faster rate of dependency length increase, this is not likely to be the result of the underlying tree structure of sentences being inherently more complex than other languages.

Regarding the **RandomFixed** baseline, we do not find that this operates differently from **RandomFree**, and intuitively this would be explained by the outputs of all random grammars being pooled together; with the resulting data being not much different to what we would see if we simply randomized all sentences. This can and should be fixed in future research.

The **FittedGrammar** baseline is more chaotic than we anticipated. In most languages towards the right of the graph, we see a small gap between the original sentences and the FittedGrammar output, though in some languages this gap is greater than in others. Many of these seem to be languages with flexible word order, such as as the Baltic languages and Greek, but also, for some reason, Danish. This could be cautious evidence of languages using their available word order flexibility to reduce dependency lengths.

However, as we reach the SOV and mixed languages on the left side of the graph, the picture is more incoherent. In Hindi and Urdu, the fitted grammar results in a higher dependency length increase even than both random baselines. We are unsure how to interpret this, and further linguistic analysis of the permutations produced by the fitted grammar is in order.

4.2 Intevener Complexity Measure

Fig. 6 shows the coefficients of the linear mixed-effects model, this time using Intervener Complexity Measure of each sentence as the response variable.

As with dependency length, we find a clear pattern whereby SOV languages, or languages with frequent verb-final constructions, show a faster rate of increase in ICM compared with SVO languages. For other languages, however, a very similar pattern of minimisation is observed, though in this case the gap between coefficients is much smaller.

Welsh once again shows the slowest rate of increase, though in this case the effect is less pronounced. Again, Irish is not among the languages with the lowest coefficients, which indicates that this is probably not due to typological properties of VSO or Celtic languages.

The observed ICM is almost colinear with OptimalOrder for several of the languages (mainly those with SVO word order), and in some cases is lower. The OptimalOrder algorithm was developed to minimise dependency lengths, not ICM, so this is unlikely to represent the true optimum. However, this finding is compelling because it suggests that observed sentences are close to an optimal ICM, while also being clearly separated from the random baselines.

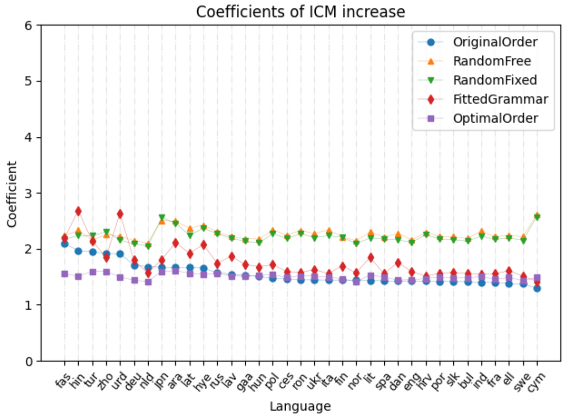

Figure 6: The coefficients of intervener complexity measure increase for all baselines in all languages. Languages are sorted in descending order by OriginalOrder coefficient.

5 Discussion

Overall, our results seem to uphold DLM as a universal, though with the ame asymmetry between verb-final and verb-initial or -medial languages. We also find this same asymmetry in intervener complexity, with the same languages showing a faster rate of increase in ICM, showing that the antilocality effect extends to this measure as well.

The next step is to turn our attention to explanations of this tendency for reduced DLM in SOV/verb-final languages. There is already work underway to explain these tendencies (Yadav et al., 2020; Jing et al., 2022).

The use of a parallel corpus has supported the results of previous research in this area. In other words, we do not see a very different picture when using a parallel corpus. An interpretation of this that dependency length and intervener complexity minimisation effects are strong enough that they show through the noise of domain and sentence variation.

However, we still maintain that parallel corpora should be used wherever possible in such studies. Our study has applied to languages as a whole, using the full range of sentences each language in the corpus. On the other hand, we hypothesise that the more focused the linguistic structures under investigation - for example, verb phrases with single object and oblique arguments (Liu, 2020), or verb phrases with two oblique arguments (Liu, 2022) - the more the noise of differing domains and sentence content will affect the results. It is particularly these kinds of studies that we believe will benefit from large parallel corpora.

A meta-study of dependency length and related experiments using both Universal Dependencies and parallel corpora would be useful to measure the extent to which such noise affects different kinds of experiments. We leave this for future research.

There are also some improvements that could be made to this study in particular.

We would like to find an algorithm for finding the linear ordering that truly optimises intervener complexity measure, so that we can properly assess how close observed orderings are to this baseline. We are unaware of such an algorithm as of yet, and Gildea and Temperley's algorithm is an imperfect stand-in. This would be particularly valuable because of the tentative evidence we find for observed word order reflecting optimised ICM.

Some previous studies have used Surface Universal Dependencies (SUD) annotated corpora (Gerdes et al., 2018) instead of Universal Dependencies. While we do not expect vastly different results, there is some contention that SUD is more appropriate for modelling syntactic difficulty and cognitive demand (Yan and Liu, 2019), and it would be beneficial to compare experiments on corpora using each of the two formalisms.

Finally, as more languages are added to CIEP+, we hope to be able to expand our analyses to more languages, particularly non-Indo-European languages.

6 Conclusion

Our replication of a keystone study on dependency length minimisation as a language universal on a much larger, parallel parsed corpus has corroborated previous findings that show evidence of systematic dependency length minimisation in a variety of the world's languages, controlling for the effect of sentence and domain variation. We find a similar effect for intervener complexity measure.

We make available our code for permuting parsed corpora according to different permutation baselines, and for analysing them in terms of dependency length, intervener complexity and other properties.[7]

We plan to use this corpus in further replications and original studies on syntactic complexity and word order constraints. Among our topics of interest are research into why dependency length minimisation is less of a pressure in verb-final languages; and the extent to which other constraints such as information locality (Futrell, 2019; Liu, 2022) and memory-surprisal tradeoff (Hahn et al., 2020, 2021) subsume dependency length as an explanatory factor for word order.

Limitations

While the design of the CIEP+ corpus is parallel in the sense that the same collection of books is to be added for each language, not all languages have the full collection. This also means that languages will have different data sizes and different book coverage. While in data exploration we did not find that the book that sentences came from was a strong random effect, it is possible that these differences may nevertheless confound the results. Book translations are continually added to the corpus, so

[7] https://github.com/andidyer/DependencyLengthSurvey

this problem will hopefully become lesser in future studies.

In contrast to the gold Universal Dependencies data used in many other studies, CIEP+ is predictively parsed, and parser error may propagate to give erroneous results. Interesting findings for any particular language should therefore be looked at with the performance of that language's Stanza model in mind.[8] CIEP+ does not currently have gold evaluation sets, so it is unfortunately not possible to get LAS scores for the models on CIEP+; we rely on the models' evaluation scores on the test sets of the UD corpora on which they are trained.

The use of a linear mixed effects model for plotting the increase in dependency length is not ideal due to the heteroscedacity of sentence dependency length relative to sentence length; variance of dependency length increases with sentence length, and means do not increase linearly. This is contrary to the assumptions of linear models, and may affect the reliability of the results. (van den Berg, 2021) We experimented with generalised mixed effects models with a Poisson link function, but found that this caused unacceptably long training times with the size of our data. We might overcome this with bootstrap sampling, or an alternative regression algorithm or software.

Ethics

We are not aware of any adverse impacts on any individual or group of individuals as a result of our study.

This paper and all associated code and statistical analysis was produced by human effort of the authors. At no point was any generative artificial intelligence used.

Our data is not available to share in its original form due to copyright concerns. However, upon request, we can provide the data in delexicalised form.

Acknowledgements

Thanks to Michael Hahn and Annemarie Verkerk for their supervision and feedback, and to Luigi Talamo and Luca Brigada Villa for comments and suggestions.

CIEP+ exists thanks to the work of Luigi Talamo and Annemarie Verkerk, and the many individuals who collected the required books. And, of course, the writers and translators of those books.

Finally, many thanks to the anonymous reviewers for their kind comments, critiques and suggestions.

[8] https://stanfordnlp.github.io/stanza/performance.html

References

Jennifer Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76:28–55.

Östen Dahl. 2007. From questionnaires to parallel corpora in typology. *Language Typology and Universals*, 60(2):172–181.

Matthew S. Dryer. 2013. Order of subject, object and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Matthew S. Dryer and Orin D. Gensler. 2013. Order of object, oblique, and verb (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Ramon Ferrer-i Cancho. 2017. The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics*, 39.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

Richard Futrell and Edward Gibson. 2015. Experiments with generative models for dependency tree linearization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1978–1983, Lisbon, Portugal. Association for Computational Linguistics.

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

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In *Translation Studies in Scandinavia*, pages 88 – 95.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*, Brussels, Belgium.

Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 94 – 126.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

Daniel Gildea and David Temperley. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 128:726–756.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 117:2347–2353.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel english–chinese dependency treebank. *Language Sciences*, 50:93–104.

Yingqi Jing, Damián Blasi, and Balthasar Bickel. 2022. Dependency length minimization and its limits: A possible role for a probabilistic version of the final-over-final condition. *Language*, 98.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Natalia Levshina and Steven Moran. 2021. Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, 7(s3):20200081.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.

Zoey Liu. 2020. Mixed evidence for crosslinguistic dependency length minimization. *STUF - Language Typology and Universals*, 73(4):605–633.

Zoey Liu. 2021. The crosslinguistic relationship between ordering flexibility and dependency length minimization: A data-driven approach. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 264–274, Online. Association for Computational Linguistics.

Zoey Liu. 2022. A multifactorial approach to crosslinguistic constituent orderings. *Linguistics Vanguard*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marius Popescu. 2011. Studying translationese at the character level. *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 634–639.

Tiina Puurtinen. 2003. Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature. *Literary and Linguistic Computing*, 18(4):389–406.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Luigi Talamo and Annemarie Verkerk. 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.

Stéphanie M. van den Berg. 2021. *Analysing Data using Linear Models*, 5 edition, chapter 7. University of Twente, Netherlands.

Stephen J. Williams. 1980. *A Welsh Grammar*. University of Wales Press, Cardiff.

Himanshu Yadav, Shubham Mittal, and Samar Husain. 2022. A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, 6:147–168.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science*, 44(4):e12822.

Jianwei Yan and Haitao Liu. 2019. Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand? In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 16–24, Paris, France. Association for Computational Linguistics.

Xiang Yu, Agnieszka Falenska, and Jonas Kuhn. 2019. Dependency length minimization vs. word order constraints: An empirical study on 55 treebanks. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 89–97, Paris, France. Association for Computational Linguistics.