Does One Size Fit all in Crosslinguistic Dependency Length Minimization?

Zoey Liu
Boston College
zoey.liu@bc.edu

Ria Upreti
University of Texas at Austin
upreti.ria@gmail.com

Mathew A. Kramer
University of Michigan
arkram@umich.edu

Savithry Namboodiripad
University of Michigan
savithry@umich.edu

Abstract

Previous studies have claimed that language structures tend to minimize the linear distance between syntactic heads and their dependents, a principle known as Dependency Length Minimization (DLM). These studies, however, have largely focused on written modality. In this study we examine the role of dependency length in acceptability ratings of English and Hindi, two typologically distinct languages, using *audio stimuli*. With double PP constructions as a test case, our results demonstrate no effect of DLM, suggesting the preference for shorter dependencies is different in acceptability and written texts. These findings are further supported with corpus analysis of a total of 10 treebanks for the two languages, which shows additional language-specific differences in the extent of DLM. We discuss the implications of our work and call for more careful consideration of linguistic and modality-specific diversity when it comes to processing-based claims about language typology.

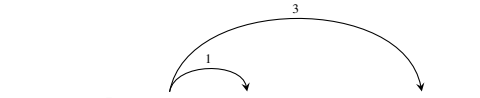
Keywords: dependency length minimization; acceptability judgment; syntactic typology

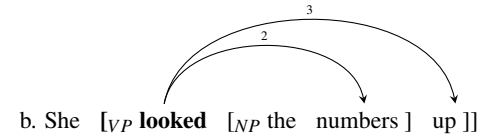
Motivation

Initially inspired by studies on language comprehension (Gibson, 1998), the principle of Dependency Length Minimization (DLM; Ferrer-i Cancho, 2004) predicts that words or phrases that are syntactically dependent on each other will tend to occur closer together, therefore minimizing the overall dependency length of the sentence. This tendency has been said to be motivated by communicative efficiency (Hawkins, 1994, 2004), such that constituent orders of shorter dependencies are preferred in order to reduce structural complexity and lessen processing difficulty (Temperley & Gildea, 2018).

As an illustration of how DLM applies to ordering preferences, consider the following examples of verb-particle constructions in English, a predominantly SVO language.

(1)





Here the verb phrase (VP) in each sentence has the same two dependents, the particle *up* and the noun phrase (NP) *the numbers*. Switching the order of these two dependents leads to two grammatical alternatives, the meanings of which are still largely comparable (Bresnan, Cueni, Nikitina, & Baayen, 2007). However, by placing the shorter constituent, *up*, closer to the head verb *looked*, the overall dependency length of (1a) is shortened compared to (1b). DLM would therefore predict that the structure of (1a) is preferable to that of (1b).

Numerous studies have explored the role of dependency length in syntactic ordering preferences (Gildea & Temperley, 2010; H. Liu, 2008; Temperley, 2007; Wasow, 1997; Wasow & Arnold, 2003; Yamashita & Chang, 2001). Corpus work looking at crosslinguistic patterns of DLM has claimed that the preference for shorter dependencies is language-universal. Using data from 37 languages, Futrell, Mahowald, and Gibson (2015) demonstrated that the observed dependency length of a given sentence tends to be shorter than when the dependency structures of the sentence are randomized. Rather than making comparisons with randomized dependency trees, Z. Liu (2020) took data from 34 languages and examined specific syntactic constructions that allow for grammatical alternatives. Through an examination of adpositional phrase (PP) constructions, which permit flexible constituent orderings (e.g., *she* ***walked*** *[$_{PP_1}$* ***with*** *friends] [$_{PP_2}$* ***for*** *an hour]* vs. *she* ***walked*** *[$_{PP_1}$* ***for*** *an hour] [$_{PP_2}$* ***with*** *friends]*), her results demonstrated a typological tendency for DLM in syntactic alternations.

As fruitful as prior experiments have been, they have mainly focused on *written texts*. Recent work, however, has shown that **the extent of DLM may vary when it comes to different modalities**. Comparing conversational speech from the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992) to written texts from the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993) in English, Z. Liu (2019) found the preference for shorter dependencies is weaker in the spoken domain. On the other hand, Kramer (2021) contrasted naturalistic speech in seven languages collected from YouTube channels to written data from the Universal Dependencies project (version 2.6; Zeman et al., 2020). His results demonstrated a higher degree of DLM in speech than in writing in the head-final languages, yet the opposite patterns hold in head-initial languages. These observations indicate that even within the same modality, there could be language-specific differences in the preference for DLM.

The mixed evidence discussed above calls for more explo-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ration of how dependency length affects word orders across registers of typologically distinct languages. That being said, there has been very little work devoted to this topic. Aside from examinations of crosslinguistic data from spoken corpora (Kramer, 2021) or comprehension studies (Gibson, 1998), a few studies have conducted online production experiments in languages other than English. For instance, Yamashita and Chang (2001) showed that there is a "long-before-short" preference in preverbal contexts of Japanese, where shorter constituents tend to be placed closer to the head verb, thus adhering to DLM. Also looking at online production, Faghiri and Samvelian (2020) found strong preferences for shorter dependencies in Persian. Both studies have investigated transitive and/or ditransitive structures.

This paper contributes to the aforementioned gaps in previous work. Rather than concentrating solely on written data or online comprehension/production, we examine the modality of *acceptability judgments* across two typologically distinct languages, English and Hindi (a predominantly head-final Indic language). With the double PP construction as the test case, our preregistered study[1] asks whether DLM influences sentence acceptability ratings.

Though most acceptability experiments used written stimuli, we used **audio stimuli** (Ferreira & Swets, 2005; Namboodiripad, 2017) to address our question. While audio stimuli have long been used to investigate certain topics, such as disfluency (Ferreira & Bailey, 2004) or prosody (Fodor, 2002), combining auditory stimuli with acceptability judgment experiments is relatively recent and has received comparatively less attention (Scontras, Polinsky, Tsai, & Mai, 2017). Compared to written stimuli, audio stimuli allow one to "increase the naturalness" of acceptability studies (Beltrama & Xiang, 2016; Polinsky, 2016). This is particularly relevant when investigating understudied languages, in contrast to Written, Institutionally supported, Standardized, and Prestigious (WISPy) languages (Sedarous & Namboodiripad, 2020) (e.g., written English), which are overrepresented in psycholinguistic research (Anand, Chung, & Wagers, 2011). Participants who speak understudied languages or varieties may be more likely to understand the language in spoken or conversational contexts very well or be able to produce the language fluently, but they might not (yet) have proficiency in writing the language. Further, as most languages and varieties do not have (standardized) written forms at all, a focus on only judgments of written sentences captures a small portion of the world's linguistic diversity. Audio stimuli are one way of addressing these considerations, allowing us to be more inclusive in our experimental design.

DLM in Acceptability Judgments

The role of dependency length in sentence acceptability has been noted extensively in prior work (Francis, 2010; Goodall, 2017; Lu, Thompson, & Yoshida, 2020; Sprouse, Wagers, & Phillips, 2012),with most studies focusing on cases of *island constraints* (Ross, 1967). The general prediction is that structures with long or non-local dependencies will result in degradation of acceptability ratings (see Goodall (2021) for a summary).

In addition to using audio stimuli, our work differs from previous studies on the relationship between dependency length and acceptability ratings in four key respects. First, previous experiments have mainly compared sentences which differ significantly in their grammaticality (e.g., ungrammatical sentences that violate certain island constraints vs. those that do not) (Sprouse et al., 2012) .[2]

Second, the majority of related studies in the literature have used fundamentally different sentence types or syntactic constructions for comparisons, for instance, contrasting sentences with *wh*-dependencies to those without (Cowart, 1997; Lu et al., 2020), or comparing stimuli with different subject clausal structures (Hofmeister, Casasanto, & Sag, 2014).

Third, as a result of potential differences in grammaticality and syntactic structures, one might end up comparing sentences where the critical regions have different syntactic heads. Consider the following examples from Cowart (1997), where __ denotes the canonical position of the fronted element, *who*. In (2a), the syntactic head of *who* is the noun *portrait*, whereas in (2b), *who* is the dependent of the head verb *sell*. (Note that even if treating *who* as the dependent of the preposition in both sentences, the two prepositions have still different syntactic heads).

(2) a. *Who* did the Duchess sell [a portrait [of __]]?
b. *Who* did the Duchess sell [a portrait [to __]]?

Fourth, despite the gaps in the literature, there are some studies which have attempted to address the three issues listed above (Francis, 2010; Futrell & Levy, 2019). These studies examined the effect of dependency length on acceptability judgments for grammatical alternatives of the same syntactic constructions, such as relative clause extraposition (Francis, 2010), and verb-particle structures (Futrell & Levy, 2019). In these cases, the constituents whose orders were switched still had the same syntactic head. Though the aforementioned experiments on syntactic alternations bear direct resemblance to our study, each investigated only English. Meanwhile, crosslinguistic comparisons, especially to languages that are typologically distinct from English, are lacking in general (Chacón, 2021).

Experiments

Using audio stimuli, we ask whether and to what extent dependency length predicts sentence acceptability ratings in English and Hindi.

Stimuli The experimental items consisted of double PP (prepositional phrase) constructions, which consist of verb

[1]The preregistration, code and data for our experiments are in quarantine at `thegoodplace.com`.

[2]Though cf. Brown, Fanselow, Hall, and Kliegl (2021) and Goodall (2015).

Table 1: Sample stimuli in English and Hindi. RCs in English and APs in Hindi are italicized and underlined; the Hindi example in the LONG-LONG condition means: *Aunt **clapped** [$_{PP_1}$ **under** the very old tree] [$_{PP_2}$ **for** the cute girl*].

Language	Condition	Example
English	SHORT-SHORT	The tourist **slept** [$_{PP_1}$ **on** the bed] [$_{PP_2}$ **throughout** the evening]
	LONG-SHORT	The tourist **slept** [$_{PP_1}$ **on** the bed *that was small*] [$_{PP_2}$ **throughout** the evening]
	SHORT-LONG	The tourist **slept** [$_{PP_1}$ **on** the bed] [$_{PP_2}$ **throughout** the evening *that was hot*]
	LONG-LONG	The tourist **slept** [$_{PP_1}$ **on** the bed *that was small*] [$_{PP_2}$ **throughout** the evening *that was hot*]
Hindi	SHORT-SHORT	mamiji [$_{PP_1}$ ped ke **neeche**] [$_{PP_2}$ bacchi ke **liye**] **taali baja rahi thi**
	LONG-SHORT	mamiji [$_{PP_1}$ *bahut puraane* ped ke **neeche**] [$_{PP_2}$ bacchi ke **liye**] **taali baja rahi thi**
	SHORT-LONG	mamiji [$_{PP_1}$ ped ke **neeche**] [$_{PP_2}$ *ek pyaari-si* bacchi ke **liye**] **taali baja rahi thi**
	LONG-LONG	mamiji [$_{PP_1}$ *bahut puraane* ped ke **neeche**] [$_{PP_2}$ *ek pyaari-si* bacchi ke **liye**] **taali baja rahi thi**
		*Aunt **clapped** [$_{PP_1}$ **under** the very old tree] [$_{PP_2}$ **for** the cute girl*]

phrases (VPs) with exactly two PP dependents (obliques) occurring on the same side. While English is a prepositional language, Hindi is postpositional. Based on patterns from corpus data (Z. Liu, 2020), double PP constructions in English tend to be mostly head-initial, in the sense that the two PPs occur postverbally. Hindi is the mirror-image of this, where the two PPs appear before the head verb (Table 1).

Stimuli in each language followed their canonical PP ordering patterns. Each item had an animate subject and an intransitive head verb; the head verbs were immediately adjacent to the two PP dependents. We further manipulated the overall dependency length of each item by adding modifiers to one or both PPs. Specifically, we attached relative clauses (RCs) to the nominal head of the PP for the English stimuli, and for Hindi we used adjectival phrases (APs), which are analogous. This manipulation resulted in four conditions for each item: SHORT-SHORT, LONG-SHORT, SHORT-LONG, and LONG-LONG. We created 20 lexicalization sets for English and 24 for Hindi. Each item was recorded separately, and its pitch contour was checked in *Praat* (Boersma & Weenink, 2010) to ensure that prosody was consistent within each condition. For each item, the two PPs in the SHORT-SHORT condition were always of equal length; however, the lengths of the two PPs in the SHORT-SHORT conditions could differ *across* items[3]. The same holds for the respective modifiers of the two PPs from the LONG-LONG conditions.

Procedure We recruited 128 participants who grew up hearing and/or speaking English in the United States. Participants were asked to listen to audio stimuli in English and rate how natural each sentence sounded on a 1-7 Likert scale. Each participant heard 5 items from each of the four conditions, along with 60 fillers of varying acceptability. The same process was carried out with 73 participants who grew up hearing and/or speaking Hindi, except that they listened to and rated audio stimuli in Hindi. Each participant heard 6 items from each condition, plus 69 fillers of varying acceptability[4].

[3]In Hindi, the nominal head of certain PPs contains a clitic dependent. This clitic may be treated as either an individual token or as an affix, depending on different writing preferences or annotation standards. To account for this discrepancy, we ran all regression models twice; one counted the clitic as a separate token, while the other did not. There were no observable differences in the results.

[4]Because our study was a sub-experiment for another preregistered study, the numbers of participants, sets of stimuli, and fillers in the two languages were not set to be exactly the same. In the regression analysis, we included participant and individual item as random effects to address concerns of the mismatching numbers.

The acceptability ratings from each participant were first transformed into by-subject z-scores and subjected to mixed-effect models. In the models, the dependent variable was the by-subject z-score, and the independent variables were CONDITION (treating the LONG-LONG condition as the reference level), the length of the sentence, and the lengths of the two PPs. Interaction terms were included between CONDITION and each of the other length factors; ITEM and PARTICIPANT were included as separate random intercepts. All models were implemented in the *lme*4 package in the programming language R (Bates, Mächler, Bolker, & Walker, 2015).

$$\text{Acceptability} \sim \text{CONDITION} * (\text{SENT\_LEN} + PP_1\text{\_LEN} + PP_2\text{\_LEN} + (1|\text{PARTICIPANT}) + (1|\text{ITEM})$$

Predictions As stated in the previous section, while there have been studies relevant to ours (Francis, 2010; Futrell & Levy, 2019), the differences in items and experimental design are significant enough that we did not feel comfortable making precise predictions based entirely on prior work. Instead, we made (tentative) predictions based on existing findings of crosslinguistic DLM and acceptability judgments.

For English, given that the SHORT-SHORT condition has the shortest sentence length and also the shortest overall dependency length, it is likely to be the most acceptable and, correspondingly, to have the largest coefficient value in our mixed-effects model. Conversely, the LONG-LONG condition should have the lowest acceptability ratings and coefficient value (Häussler, Grant, Fanselow, & Frazier, 2015).

Between the other two conditions, the LONG-SHORT condition is expected to have the second-lowest ratings for two reasons: (1) It has a shorter overall sentence length than the LONG-LONG condition and thus is likely to be rated as more acceptable than these sentences; (2) By placing the PP of shorter length farther away from the head verb, the LONG-SHORT condition in English does not abide by the principle of DLM. The SHORT-LONG condition, which is consistent with DLM, is thus expected to have higher ratings.

With Hindi, for the same reasoning, the SHORT-SHORT condition is expected to have the highest acceptability ratings and, accordingly, the highest coefficient value in the regression model, while the LONG-LONG condition should be the

least acceptable. The other two conditions, as above, are expected to have higher ratings than the LONG-LONG condition due to their shorter overall length.

Nevertheless, it is not clear whether items in the LONG-SHORT condition would be more acceptable than those in the SHORT-LONG conditions, even though the former demonstrates DLM by placing the shorter PP adjacent to the head verb. Given results from Z. Liu (2020), there did not appear to be a tendency for shorter dependency lengths in at least written texts in Hindi. Across the languages examined, it seems that the preference for DLM is generally weaker in head-final contexts, where the two PPs occur preverbally. Hence it is possible that in Hindi, the acceptability ratings of sentences from the LONG-SHORT condition and those from the SHORT-LONG condition are comparable, and (much) lower than those of the SHORT-SHORT condition.

Results

As shown in Figure 1, in English, the SHORT-SHORT condition had the highest acceptability rating, while the LONG-LONG condition had the lowest. In contrast to our initial expectations, however, sentences from the SHORT-LONG condition were much less acceptable on average than the SHORT-SHORT ones; their average acceptability rating was comparable to that of sentences from the LONG-SHORT condition. This means that even when sentences abide by DLM, it does not seem to give them any "advantage" in their acceptability, at least in the context that we are investigating. Overall, the average acceptability of the four conditions was roughly correlated with their sentence lengths.

These results are further corroborated by the coefficient values of all the conditions derived from the mixed-effects models, presented in Figure 3a. Again, the SHORT-SHORT condition had the largest coefficient while the LONG-LONG condition had the lowest. The coefficients of LONG-SHORT sentences and of SHORT-LONG ones laid in between, yet there was no significant difference between the two conditions.

For Hindi, on the other hand, the SHORT-LONG condition turned out to be the most acceptable, with the highest average acceptability rating. The LONG-SHORT condition, which follows the preference for DLM, was the least acceptable. That being said, there do not seem to be strong differences in acceptability ratings between the four conditions; the largest z-score difference was 0.09, between SHORT-LONG and LONG-SHORT conditions. This observation is also reflected in Figure 3b, which does not show any significant differences in the coefficient values for the four conditions.

To further investigate whether there are systematic differences between the results for the two languages, we performed an additional regression analysis. The mixed-effect model used in this analysis was similar to the ones applied before, except that it included LANGUAGE as a fixed effect, as well as an interaction between LANGUAGE and CONDITION.

$$\text{Acceptability} \sim \text{CONDITION} * (\text{LANGUAGE} + \text{SENT\_LEN} + PP_1\text{\_LEN} + PP_2\text{\_LEN} + (1|\text{PARTICIPANT}) + (1|\text{ITEM})$$

As demonstrated in Figure 4a, when combining the acceptability ratings of the two languages together, there do not appear to be significant differences between the coefficient values of the conditions. In other words, the acceptability of sentences from different conditions is comparable regardless of whether they have shorter dependencies and/or shorter overall lengths. The acceptability of sentences in each condition does not seem to differ significantly between the two languages either (Figure 4b).

Corpus Analysis

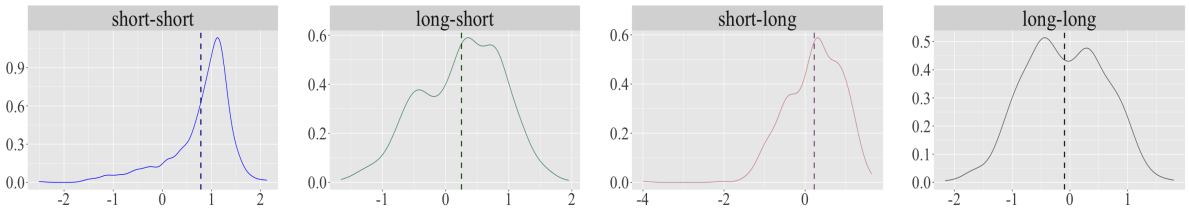
Why is the preference for shorter dependencies not reflected in our experiments? Before drawing the conclusion that dependency length does not predict PP ordering in acceptability ratings of English and Hindi, at least not with audio stimuli, we considered three other potential explanations.

First, prior work has found structural or lexical frequency to have an effect on acceptability (though to different degrees), with the least and the most frequent variants being more or the most acceptable (Bermel & Knittl, 2012; Divjak, 2008, 2017; Kempen & Harbusch, 2004). In our case, if double PP structures that adhere to DLM are *not* more frequent than those without, this would provide an alternative explanation for the results of our acceptability study.

Nevertheless, it is not realistic to estimate frequency information with the experimental settings presented here. As a coarse proxy, we turned to corpus data. For English, we took treebanks from the latest version of the Universal Dependencies project (UD) (Zeman et al., 2021); data from the learner corpus was not included. Additionally, we used the Penn Treebank (PTB) (Marcus et al., 1993) and extracted double PP structures from each of the Wall Street Journal (WSJ) (Marcus et al., 1993), the Brown corpus (Kučera & Francis, 1967), and transcriptions of spontaneous spoken conversations from the Switchboard corpus (Godfrey et al., 1992). For Hindi, we also used data from UD.

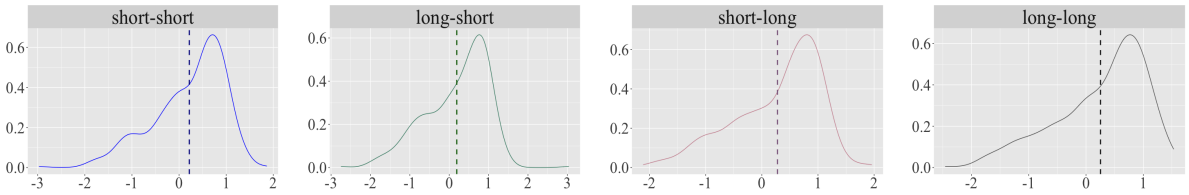
Note that although Z. Liu (2020) also analyzed PP orders from UD, the number of double PP constructions that we initially extracted is 1.14 times larger for English, and 2.54 times larger for Hindi. In addition, rather than combining the cases extracted from the treebanks of the same language together, as was done in prior work, in order to account for differences in annotation standards and data sources, we kept each treebank separate. (For the UD_English-GUM treebank, we only took data from the written domain, since only 12 double PP instances were from the spoken domain). Treebanks with fewer than 100 analyzable cases in total were excluded.

Given the double PP constructions from each treebank, we first calculated the total number of instances where the two PPs had different lengths, N. Treebanks where $N < 100$ were excluded. For each of the remaining treebanks, we calculated their DLM ratio, DLM_r. This ratio was computed as follows: First, we counted the number of cases that adhered to DLM, N_{short}, and those that did not, N_{long}. We then estimated DLM_r as $\frac{N_{short}}{N_{long}}$. Finally, we performed bootstrapping with 10,000



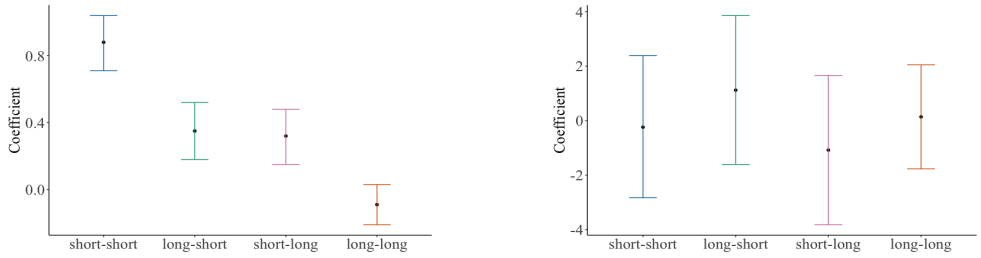
(a) mean(z) = 0.78 (b) mean(z) = 0.25 (c) mean(z) = 0.22 (d) mean(z) = -0.09

Figure 1: Density plot for by-subject z-scores of acceptability ratings for English stimuli; x-axis represents z-scores, and y-axis denotes density values.



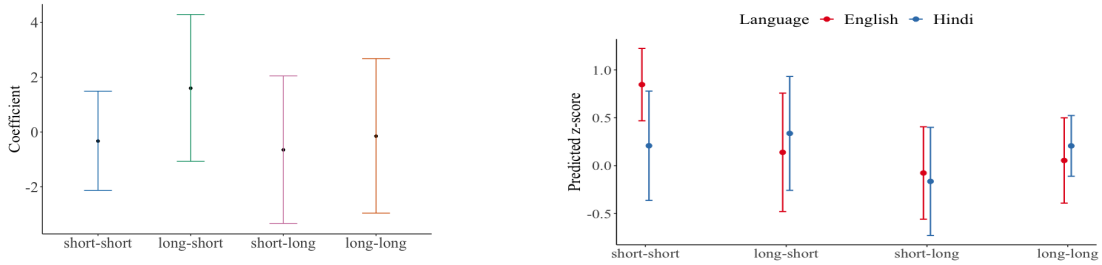
(a) mean(z) = 0.22 (b) mean(z) = 0.19 (c) mean(z) = 0.28 (d) mean(z) = 0.25

Figure 2: Density plot for by-subject z-scores of acceptability ratings for Hindi stimuli; x-axis represents z-scores, and y-axis denotes density values.



(a) English (b) Hindi

Figure 3: Coefficient values for each condition in the individual mixed-effect model for each language.



(a) Coefficient values for each condition. (b) Interaction effects between language and condition.

Figure 4: Results for regression analysis when combining data of the two languages together.

iterations for significance testing of DLM_r.

As shown in Table 2, while the preference for DLM holds in all analyzable treebanks of English in both spoken and written modalities, we found no tendency towards shorter dependencies in written Hindi. If these numbers are representative of the structural properties of the language, we would expect to see DLM reflected in acceptability judgments of the double PP structures in English, which is in opposition to what we found in our acceptability study.

Second, construction frequency could play a role in accept-

Table 2: Results for corpus analysis of double PP constructions in English and Hindi; cases where the ratio of DLM is significant is marked in bold. While both representing spoken data, UD_English-Atis contains human speech directed towards machines, while Switchboard contains naturalistic conversations.

Language	**Treebank**	**Domain**	N	DLM_r	**Overall construction (%)**	$N_{selected}$	$DLM_{r\_selected}$
English	UD_English-EWT	written	650	**2.26** (1.63, 3.25)	4.82	94	-
	UD_English-GUM		388	**2.57** (2.56, 3.88)	6.44	28	-
	UD_English-Lines		291	**3.20** (1.83, 5.93)	7.02	34	-
	UD_English-ParTut		186	**2.13** (1.16, 3.77)	10.19	12	-
	WSJ		2,755	**2.65** (2.25, 3.21)	7.31	225	**2.45** (1.45, 4.63)
	Brown		2,420	**3.47** (2.96, 4.20)	5.83	376	**2.32** (1.52, 3.70)
	UD_English-Atis	spoken	284	**3.03** (1.86, 5.92)	10.70	186	**3.20** (1.69, 7.08)
	Switchboard		883	**1.83** (1.44, 2.41)	5.80	290	**1.63** (1.07, 2.77)
Hindi	Hindi-HDTB	written	3188	1.06 (0.91, 1.22)	23.26	792	1.10 (0.85, 1.44)
	Hindi-PUD		172	1.14 (0.69, 2.07)	20.30	31	-

ability. In particular, it could be the case that the overall frequency of double PP constructions in both languages is quite high in general, leading to sentences with different dependency length conditions being comparably acceptable. To investigate this possibility further, we calculated the proportion of double PP constructions (including ones where the two PPs have equal lengths) in each treebank by dividing construction frequency by the total number of sentences in the treebank.

The results are presented as "Overall construction %" in Table 2. Double PP constructions appear to be fairly frequent in both English and Hindi; this could offer a plausible explanation for the patterns from the acceptability ratings, but would still need additional experimental support from future work. On another note, notice that the construction frequency is much higher in Hindi than that in English; it would not be unreasonable to think that these differences would be reflected in acceptability judgments, i.e., that there would be significant differences between the two languages and/or each condition, as well. This conjecture, however, is not supported by the results presented in Figure 4b.

The third potential reason pertains to the fact that our stimuli are relatively short. In English, the longest sentences have 15 words, while in Hindi the maximum length is 18. Previous work has demonstrated a positive correlation between the extent of DLM and sentence length (Futrell, Levy, & Gibson, 2020), meaning that the longer the sentence is in terms of written words, the stronger the preference for shorter dependencies. In addition, while some studies have found that DLM also holds for shorter constituents (e.g., adjectival phrases (Gulordava & Merlo, 2015)), others have argued for the opposite effect (Ferrer-i Cancho & Gómez-Rodríguez, 2021). Here the length differences between the two PPs in the LONG-SHORT and SHORT-LONG conditions were small (3 words maximum in English; 4 in Hindi). However, in his examination of PP orders in written English, Hawkins (1999) found that a bigger difference in length between the two PPs resulted in a stronger tendency towards DLM.

To examine the aforementioned issues, we selected from each treebank of English all instances in which the sentence length *and* the length difference between the two PPs equaled or were smaller than those of our English stimuli. The same process was carried out for Hindi. Treebanks where the number of the selected instances $N_{r\_selected}$ was smaller than 100 were not included. As shown in Table 2, DLM holds in English for the selected double PP structure, but does not appear to hold in Hindi. These observations suggest that the lengths of the stimuli and PPs used in our experiments cannot explain the lack of an effect of DLM in the acceptability ratings.

Conclusion

Taking double PP constructions as a test case, our study has probed the role of dependency length in acceptability judgments of English and Hindi using audio stimuli. The results have demonstrated no effect of DLM, and no significant differences between these two typologically distinct languages. On a larger scale, leveraging corpora, we concluded that there are modality-specific differences in the preference for shorter dependencies. In addition, while in English there were discrepancies in the effect of DLM between acceptability judgments and corpora, no such discrepancies were found in Hindi, indicating language-specific differences in the extent to which shorter dependencies are preferred.

Overall, our results suggest that "one size does not fit all" in crosslinguistic DLM when it comes to both languages and modalities. We hope our findings have implications for current progress in language typology, and call for researchers to be more mindful about language- or modality-specific differences when drawing conclusions from corpus-based analysis. While there have been notable exceptions examining structural variations using data beyond just the written forms (see Schnell and Schiborr (2022) for a summary), this is yet to be the common practice. In the case of DLM, we believe that one cannot fully understand its predictive role in ordering preferences until we have examined its effect in much wider contexts. While this paper examines only English and Hindi, in future work, we plan to carry out similar analyses using audio stimuli in other languages. In addition, we hope to extend this approach beyond double PP constructions in order to investigate whether there may be construction-specific effects in the apparent preference for shorter dependencies.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Anand, P., Chung, S., & Wagers, M. (2011). Widening the net: Challenges for gathering linguistic data in the digital age. *NSF SBE 2020. Rebuilding the mosaic: Future research in the social, behavioral and economic sciences at the National Science Foundation in the next decade*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beltrama, A., & Xiang, M. (2016). Unacceptable but comprehensible: the facilitation effect of resumptive pronouns. *Glossa*, *1*(1), 1.

Bermel, N., & Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in czech. *Corpus Linguistics and Linguistic Theory*, *8*(2), 241–275.

Boersma, P., & Weenink, D. (2010). *Praat: Doing phonetics by computer* [Miscellaneous]. 5.1.44. Retrieved from http://www.praat.org/

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation* (pp. 69–94). KNAW.

Brown, J., Fanselow, G., Hall, R., & Kliegl, R. (2021). Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *Plos one*, *16*(5), e0251280.

Chacón, D. A. (2021). Acceptability (and other) experiments for studying comparative syntax. In G. Goodall (Ed.), *The cambridge handbook of experimental syntax* (p. 181–208). Cambridge University Press. doi: 10.1017/9781108569620.008

Cowart, W. (1997). *Experimental syntax: applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Divjak, D. (2008). On (in) frequency and (un) acceptability. *Corpus linguistics, computer tools and applications–state of the art*, 213–233.

Divjak, D. (2017). The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in polish. *Cognitive Science*, *41*(2), 354–382.

Faghiri, P., & Samvelian, P. (2020). Word order preferences and the effect of phrasal length in SOV languages: evidence from sentence production in Persian. *Glossa: a journal of general linguistics*.

Ferreira, F., & Bailey, K. G. (2004). Disfluencies and human language comprehension. *Trends in cognitive sciences*, *8*(5), 231–237.

Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause "island" contexts. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (p. 263-278). Mahway, New Jersey: Erlbaum.

Ferrer-i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*(5), 056135.

Ferrer-i Cancho, R., & Gómez-Rodríguez, C. (2021). Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, *28*(1), 50–76.

Fodor, J. D. (2002). Psycholinguistics cannot escape prosody. In *Speech prosody 2002, international conference.*

Francis, E. J. (2010). Grammatical weight and relative clause extraposition in English. *Cognitive Linguistics*, *21*(1), 35–74.

Futrell, R., & Levy, R. P. (2019). Do RNNs learn human-like abstract word order preferences? In *Proceedings of the society for computation in linguistics (SCiL) 2019* (pp. 50–59). Retrieved from https://aclanthology.org/W19-0106 doi: 10.7275/jb34-9986

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.

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.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1 - 76.

Gildea, D., & Temperley, D. (2010). Do Grammars Minimize Dependency Length? *Cognitive Science*, *34*(2), 286-310.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 517–520).

Goodall, G. (2015). The d-linking effect on extraction from islands and non-islands. *Frontiers in psychology*, *5*, 1493.

Goodall, G. (2017). Referentiality and resumption in wh-dependencies. *Asking the right questions: Essays in honor of Sandra Chung*, 65–80.

Goodall, G. (2021). Sentence acceptability experiments: What, how, and why. In G. Goodall (Ed.), *The cambridge handbook of experimental syntax* (p. 7–38). Cambridge University Press. doi: 10.1017/9781108569620.002

Gulordava, K., & Merlo, P. (2015). Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 247–257).

Häussler, J., Grant, M., Fanselow, G., & Frazier, L. (2015).

Superiority in English and German: Cross-Language Grammatical Differences? *Syntax*, *18*(3), 235–265.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency* (Vol. 73). Cambridge: Cambridge University Press.

Hawkins, J. A. (1999). The relative order of prepositional phrases in English: Going beyond Manner–Place–Time. *Language Variation and Change*, *11*(3), 231–266.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2014). Processing effects in linguistic judgment data:(super-) additivity and reading span scores. *Language and Cognition*, *6*(1), 111–145.

Kempen, G., & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann & C. Habel (Eds.), *Multidisciplinary approaches to language production* (Vol. 157, pp. 173–182). Mouton de Gruyter.

Kramer, A. (2021). Dependency Lengths in Speech and Writing: A Cross-Linguistic Comparison via Youdepp, a Pipeline for Scraping and Parsing Youtube Captions. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 359–365.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Liu, Z. (2019, August). A comparative corpus analysis of PP ordering in English and Chinese. In *Proceedings of the first workshop on quantitative syntax (quasy, syntaxfest 2019)* (pp. 33–45). Paris, France: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-7905 doi: 10.18653/v1/W19-7905

Liu, Z. (2020). Mixed evidence for crosslinguistic dependency length minimization. *STUF - Language Typology and Universals*, *73*(4), 605 - 633. Retrieved from https://www.degruyter.com/view/journals/stuf/73/4/article-p605.xml doi: https://doi.org/10.1515/stuf-2020-1020

Lu, J., Thompson, C. K., & Yoshida, M. (2020). Chinese wh-in-situ and islands: A formal judgment study. *Linguistic Inquiry*, *51*(3), 611–623.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Namboodiripad, S. (2017). *An experimental approach to variation and variability in constituent order*. Unpublished doctoral dissertation, University of California, San Diego.

Polinsky, M. (2016). Structure vs. use in heritage language. *Linguistics Vanguard*, *2*(1).

Ross, R. J. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT.

Schnell, S., & Schiborr, N. N. (2022). Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics*, *8*, 171–191.

Scontras, G., Polinsky, M., Tsai, C.-Y. E., & Mai, K. (2017). Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A journal of general linguistics*, *2*(1), 1–28.

Sedarous, Y., & Namboodiripad, S. (2020). Using audio stimuli in acceptability judgment experiments. *Language and Linguistics Compass*, *14*(8), e12377.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 82–123.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, *105*(2), 300–333.

Temperley, D., & Gildea, D. (2018). Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal? *Annual Review of Linguistics*, *4*(1), 67-80.

Wasow, T. (1997). End-Weight from the Speaker's Perspective. *Journal of Psycholinguistic Research*(3), 347–361.

Wasow, T., & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of Grammatical Variation in English*, *43*, 119–154.

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*(2), B45–B55.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., ... Ziane, R. (2021). *Universal dependencies 2.9*. Retrieved from http://hdl.handle.net/11234/1-4611 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Agić, Ž., ... Zhuravleva, A. (2020). *Universal dependencies 2.6*. Retrieved from http://hdl.handle.net/11234/1-3226 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)