Proceedings of the Society for Computation in Linguistics

2019

Quantifying Structural and Lexical Constraints in PP Ordering Typology

Zoey Liu
University of California, Davis, yiliu@ucdavis.edu

Quantifying Structural and Lexical Constraints in PP Ordering Typology

Zoey Liu
Department of Linguistics
University of California, Davis
One Shields Avenue, Davis, CA
`yiliu@ucdavis.edu`

1 Introduction

Recent research has demonstrated across languages that the overall or average dependency lengths tend to be minimized by their grammars as a whole (Futrell et al., 2015). Other experiments looking at specific syntactic constructions of individual languages have shown that if the constituents in a sentence have flexible orderings, there is a preference for shorter constituents to appear closer to their syntactic heads (Jaeger and Norcliffe, 2009). Nevertheless, whether there is crosslinguistic preference for dependency length minimization (DLM) in specific syntactic structures with alternative constituent orderings remains unclear. Additionally, previous studies have shown contrary evidence and suggested that whether DLM exists relies on the structural characteristics of the language and the particular ordering constructions (Lohmann and Takada, 2014). Besides, what other linguistic factors can serve as effective typological determinants of word order preferences and their relationship with dependency length await further experimentation.

This study starts to bridge the gap by investigating prepositional and postpositional phrase (PP) order typology. Previous experiments regarding PP ordering have mainly focused on English (Hawkins, 1999; Liu and Sagae, 2018). Using multilingual corpora from the Universal Dependencies project (Nivre et al., 2017), we present the first large-scale crosslinguistic exploration of PP orderings across 31 languages. More specifically, we focus on sentences with verb phrases (VP) that contain exactly two PPs, the ordering of which in some contexts allows certain flexibility. In other words, the placement of the two PPs is either less or not constrained by the grammar of the language and hence language users have a choice towards the relative ordering of the PPs. We examine the effects of four constraints that have been suggested to affect constituent orderings and/or language processing: *dependency length*, *semantic closeness*, *lexical frequency* and *word co-occurrence information*.

2 Related Work & Motivation

Dependency length Previous studies have shown that PP orderings in English demonstrate DLM where shorter PP dependents tend to appear closer to the head verb. We investigate whether such tendency holds for PP ordering typology. To estimate the effect of dependency length, for each instance we measured and compared the length[1] of the PP closer to the verb and of the PP farther from the verb.

Semantic closeness Though it has been suggested constituents that are semantically related should occur together syntactically (Wasow and Arnold, 2003), the precise meaning of semantic closeness as well as the ideal way to compute it have not been explored much. We investigate whether the PP which is semantically closer to the head verb would appear closer. We approximate semantic closeness as how semantically similar the head verb and the nominal head of each PP are, measured with cosine similarity and word embeddings[2]. We consider the PP with a nominal head that is semantically more similar with its head verb to be the PP that is semantically closer.

Lexical frequency Few studies have looked at the role of lexical frequency in syntactic orderings. Morgan and Levy (2015) demonstrated that the more frequent word tends to appear first in binomial expressions. Liu and Sagae (2018) presented similar effects for lexical frequency in PP orders in English. Following Liu and Sagae (2018), we

---

[1] The length of each PP is measured as the number of tokens it has following the annotation scheme of the corpora.
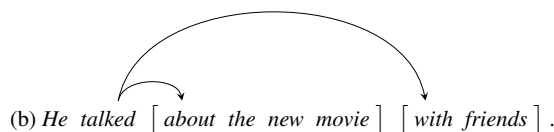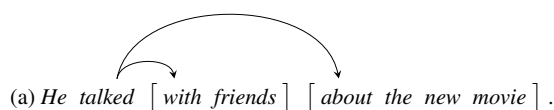[2] `https://github.com/bheinzerling/bpemb`

investigate whether there is crosslinguistic preference for the PP with more frequent words to occur first. We approximate lexical frequency of each PP with estimations from language unigram model for each language using a Wikipedia dump[3].

Word Co-occurrence Information Recent studies indicate that words that are more likely to co-occur with each other prefer to appear closer (Futrell and Levy, 2017). We investigate whether the PP that is more likely to co-occur with the head verb would be placed closer. We use pointwise mutual information (PMI) to measure the co-occurrence information of the head verb and the nominal head within each PP. We regard the PP with a nominal head that has higher PMI with the head verb to be the PP that is more likely to co-occur with the verb.

3 Experiments & Results

Our results[4] show four different PP orders as described in Table 1.

3.1 Languages with head-initial PP after head verb

(a) *He talked* [*with friends*] [*about the new movie*] .

(b) *He talked* [*about the new movie*] [*with friends*] .

For consistently head-initial languages like English, when a VP contains two PP dependents on the same side of the head verb, it will have head-initial PPs after the head verb, where the headedness of the PPs matches that of the VP. Based on our initial conjectures regarding the effects of the four factors, it is clear that dependency length, semantic closeness and word co-occurrence information always pull in the same direction regardless of the PP ordering structure of the language. If we assume that shorter PPs tend to be more frequent, a preference for shorter dependencies will place the shorter PP closer to the verb, which also means that the presumably more frequent PP will occur first. In this case,

lexical frequency will be co-operating with the other three factors in the same direction as well.

Eleven languages demonstrate the same PP ordering pattern as English, all favoring DLM. The number of sentences with the shorter PP closer to the verb is on average 4 times that of sentences with the longer PP closer to the verb. There does seem to be a general preference for PPs with more frequent words to appear first; nonetheless, the effect is not quite as pronounced as that of dependency length. On the other hand, semantic closeness and word-co-occurrence information do not appear to affect PP orderings overall.

3.2 Languages with head-final PP before head verb

For consistently head-final languages like Japanese, the headedness consistency between the PPs and the VP also exists, except that both are head-final. When the shorter PP is closer to the verb, which reduces overall dependency length, it is the second in the sequence of the two PPs and the longer PP appears first. However, if the shorter PP is in fact more frequent, that will violate the expectation that the more frequent PP should occur first. In this case, lexical frequency will be pulling in different directions from the other three factors. Thus we expect effects based on dependency length, semantic closeness, and word co-occurrence information to be weaker than in languages like English, where the four constraints are in agreement with regard to PP ordering.

(c) 彼は [赤い ペン で] [自分 で] 書いた
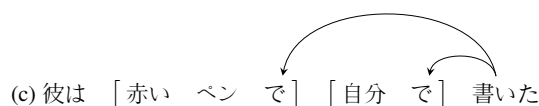
He [a red pen with] [himself by] wrote.

(d) 彼は [自分 で] [赤い ペン で] 書いた
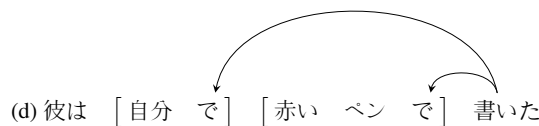
He [himself by] [a red pen with] wrote.

'He wrote by himself with a red pen.'

In our dataset, Japanese and Hindi show VP instances with exactly two head-final PP dependents before the head verb. Though there appears to be a preference for DLM in Japanese, the effect

Word order features	PP ordering pattern	Languages
Consistently head-initial	head-initial PP after head verb	English, Arabic, Danish, Galician, Greek, Hebrew, Indonesian, Norwegian, Serbian, Slovak, Swedish
Consistently head-final	head-final PP before head verb	Japanese, Hindi
Relatively free word order	head-initial PP before or after head verb	German, Bulgarian, Croatian, Czech, Dutch Latin, Russian, Slovenian, Ukranian Catalan, French, Italian, Portuguese, Romanian, Spanish Afrikaans, Persian
Typologically mixed	head-initial PP before head verb	Chinese

Table 1: PP ordering patterns of four types of languages based on our dataset.

of dependency length is much weaker compared to consistently head-initial languages. The ratio between the number of VPs with the shorter PP closer to the verb and the number of VPs with the longer PP closer to the verb is 1.7, which is substantially lower than the 4.0 average ratio for languages as English, and even substantially lower than the lowest of these ratios, 2.4. The contrast between Hindi and consistently head-initial languages is even more pronounced, with a slight overall preference (1.14 ratio) for *longer* PPs to be closer to the verb. No effects have been observed for either semantic closeness or word co-occurrence information. Despite this, lexical frequency seems to play a role for PP ordering in Japanese. The overall patterns for the four factors speak to our initial conjecture that in consistently head-final languages, lexical frequency is partially competing against the other three factors.

3.3 Languages with PPs before or after head verb

(e) *dass er* [*auf dem Tisch*] [*mit ihr*] *isst*

that he [on the table] [with her] eats

(f) *dass er* [*mit ihr*] [*auf dem Tisch*] *isst*

that he [with her] [on the table] eats

'that he eats with her on the table.'

Whether the PPs and the VP are consistently head-initial or head-final, the English and Japanese examples demonstrate that having matching headedness between the VP and the PPs

results in optimal overall dependency length when the shorter PP is closer to the verb. However, for some languages with relatively free word order, particularly those with a mix of verb-initial and verb-final clausal structures such as German or Czech, the PP ordering pattern is more complicated. These languages have predominantly head-initial PPs, which show the same orderings as English. Besides, they also exhibit cases where two head-initial PPs appear before the head verb, as in (e) and (f). However, when head-initial PPs appear before the verb, the cost of the longest dependency between the head verb and the first PP is incurred no matter what the order of the PPs is. Hence in languages with inconsistent headedness between PPs and VPs, there may not be an as pronounced preference for DLM. Nevertheless, since semantic closeness and word co-occurrence information are less subject to the cost of the long dependency, it is possible that these two factors are able to comparatively account for more of PP ordering patterns in these languages. Similar to Japanese, since the two PPs appear before the head verb, lexical frequency will be working in opposition to the other three factors as well.

Fifteen languages in our dataset have head-initial PPs appearing both before or after the head verb. When the PPs appear after the head verb, ordering preferences for DLM are closely aligned with those that are consistently head-initial. When the PPs appear before the head verb, however, the observations for DLM are substantially weaker. The six Romance languages in particular (Catalan, French, Italian, Portuguese, Romanian, Spanish) exhibit uniform patterns that are against DLM, where longer PPs appear closer to the head verb. Lexical frequency appears to play a role when the PPs are after the head verb, with results comparable to those for the consistently head-initial lan-

guages. Whereas we observe mixed patterns when the PPs occur before the verb as well. No significant effects have been found for semantic closeness and word co-occurrence information regardless of where the PPs appear.

Besides the fifteen, two other languages with relatively free word order, Afrikaans and Persian, present only cases where two head-initial PPs occur before the head verb. Both languages show a preference for shorter dependencies, yet both present the opposite ordering observations for lexical frequency. Semantic closeness does not seem to be effective. On the other hand, word co-occurrence information seems to play the strongest role among the four factors for the two languages.

3.4 Languages with head-initial PP before head verb

(g) 这些 [和 他 的 预言] [在 表面] 有 出入
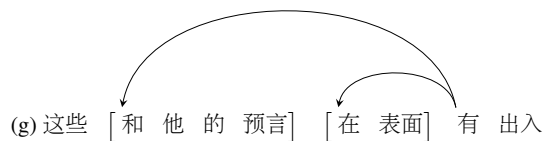
These (comments) [with his prophecy] [on the surface] have differences.

(h) 这些 [在 表面] [和 他 的 预言] 有 出入
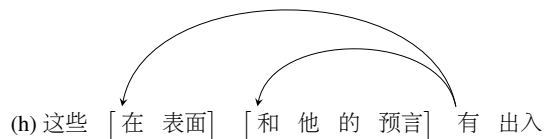
These (comments) [on the surface] [with his prophecy] have differences.

'These comments have differences on the surface with Virgil's prophecy.'

For typologically mixed languages as Mandarin Chinese, the inconsistency of headedness between the PPs and the VP is also observed. In the Chinese data, all found VP instances have head-initial PPs before the head verb. While the PP orders present opposite preferences for DLM, the other three factors show positive effects. This suggests to our initial conjecture that when inconsistent headedness exists, that when the long dependency is already incurred, the other three factors may potentially have more pronounced effects.

In sum, we demonstrate empirically that the four factors are co-operating as well as competing

motivations for PP orderings, and that the effects of these constraints are not necessarily language-universal, but dependent on the structural features of different language types. The fact there is certain amount of data where none of the four constraints are effective indicates there are other linguistic motivations, possibly involving discourse structure, that come into play. We leave that for future work.

References

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. page 688–698.

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.

John A. Hawkins. 1999. The relative order of prepositional phrases in english: Going beyond manner–place–time. *Language variation and change* 11(3):231–266.

T. Florian Jaeger and Elisabeth J. Norcliffe. 2009. The cross-linguistic study of sentence production. *Language and Linguistics Compass* 3(4):866–887.

Zoey Liu and Kenji Sagae. 2018. Dependency length minimization and lexical frequency in prepositional phrase ordering in english. *Proceedings of the Society for Computation in Linguistics* 1(1):189–192.

Arne Lohmann and Tayo Takada. 2014. Order in np conjuncts in spoken english and japanese. *Lingua* 152:48–64.

Emily Morgan and Roger Levy. 2015. Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. page 1649–1654.

Joakim Nivre, Željko Agić, Lars Ahrenberg, and et al. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2515.

Thomas Wasow and Jennifer Arnold. 2003. Post-verbal constituent ordering in english. *Topics in English Linguistics* 43:119–154.