Multiple center-embedding is more common in verb-final languages

Multiple center-embedding of clauses (as in “the rat the cat the dog worried chased ate the cheese”) is commonly assumed, on the basis of observational and experimental evidence, to present difficulties for processing (Chomsky & Miller 1963, Gibson & Thomas 1998) and therefore to be disfavored in actual language use. Furthermore, corpus research has found that center-embedding occurs very rarely in writing in European languages including German, French, and English (Karlsson 2007) and is nearly nonexistent in spontaneous speech (Karlsson 2020 on English; Laury & Ono 2010 on Japanese and Finnish). However, these corpus studies are generally limited to European languages with predominantly SVO order, and there is little data, particularly from written corpora, in languages with stricter SOV order such as Japanese and Korean, in which center-embedding may be more prevalent. The current study presents results from a typologically broader survey of languages, and demonstrates that the prevalence of multiple clausal center-embedding is greater in verb-final languages.

We hypothesized that verb-final (SOV basic order) languages are more tolerant of multiple center-embedding than verb-medial (SVO) languages (or those with mixed headedness or free word order), because their word order forces center-embedding of clauses in certain cases. For example, complement clauses in verb-final languages will be center-embedded, intervening between the subject and the main clause verb; in SVO languages like English, these clauses are right-branching (Kuno 1974). However, SOV languages may avoid center-embedding by extraposition of clauses (Dryer 1980) or via scrambling (Lewis & Nakayama 2001). If these strategies are frequently used because center-embedding presents a particular processing difficulty, it would be expected that actual rates of multiple center-embedding would be consistently low across languages, regardless of word order. In addition, recent research on dependency length minimization (Ying et al. 2021) suggests that SOV languages have a weaker tendency to minimize dependency lengths, which has implications for the prevalence of center-embedding. Since center-embedding produces greater dependency lengths (Hawkins 2004) and may be disfavored for minimization reasons, it seems likely that languages that are more tolerant of long dependencies will be more permissive of clausal center-embedding specifically.

To test this hypothesis regarding word order and multiple center-embedding, corpus searches were performed. Selected corpora in the Universal Dependencies database (Nivre et al. 2018) were searched using a pattern of dependency tagging that specified the existence of (at least) three subject-verb dependencies, in nested order. This pattern thus represented a main clause with two center-embedded clauses. All available corpora for the following languages were searched: the SVO languages English, French, Spanish, Finnish, Swedish, Danish, and Italian; the mixed-headedness languages Dutch and German; and the SOV languages Japanese, Korean, Latin, Persian, and Turkish. Altogether these corpora contained over 700,000 sentences. Matched sentences were then examined to remove false positives (e.g. incorrect syntactic tagging, or parenthetical clauses that were not syntactically integrated into the higher clause), with the assistance of fluent speaker linguists (some of the Japanese and Turkish sentences still need to be checked at the time of this writing).

There is a general correlation between basic word order and prevalence of center-embedding: multiple center embeddings, as defined in the search described above, are much more common in SOV languages than in other types. This can be observed clearly in Figure 1, which displays the prevalence of center-embedding (matching sentences per 10,000) across the language corpora that were searched. There is a very clear difference between SOV languages and SVO/mixed-headedness languages, confirmed statistically by Fisher’s exact test ($p < .0001$). While overall rates of center-embedding are still low even in the most embedding-rich language, Korean (~1 in 1400), this is still much higher than the prevalence observed in languages with SVO order, where center-embedding was very rare (~1 in 30,000) or nonexistent in the searched corpora. Persian is also unusual among verb-final languages in its low rate of center-embedding, possibly due to specific rules governing obligatory extraposition (Dryer 1980).

These results challenge the widespread assumption that multiple center-embedding is generally disfavored, and suggest that the prevalence of this construction is in fact contingent on language-specific syntactic factors rather than entirely on language-independent processing factors. They also dovetail with earlier experimental work suggesting that the ability to process center-embedded sentences correctly also varies across languages (e.g. Vasishth et al. 2010, Frank et al. 2019), and highlights the importance of analyzing both experimental and naturalistic data from a variety of languages to support theories about the processing of uncommon and difficult syntactic constructions.

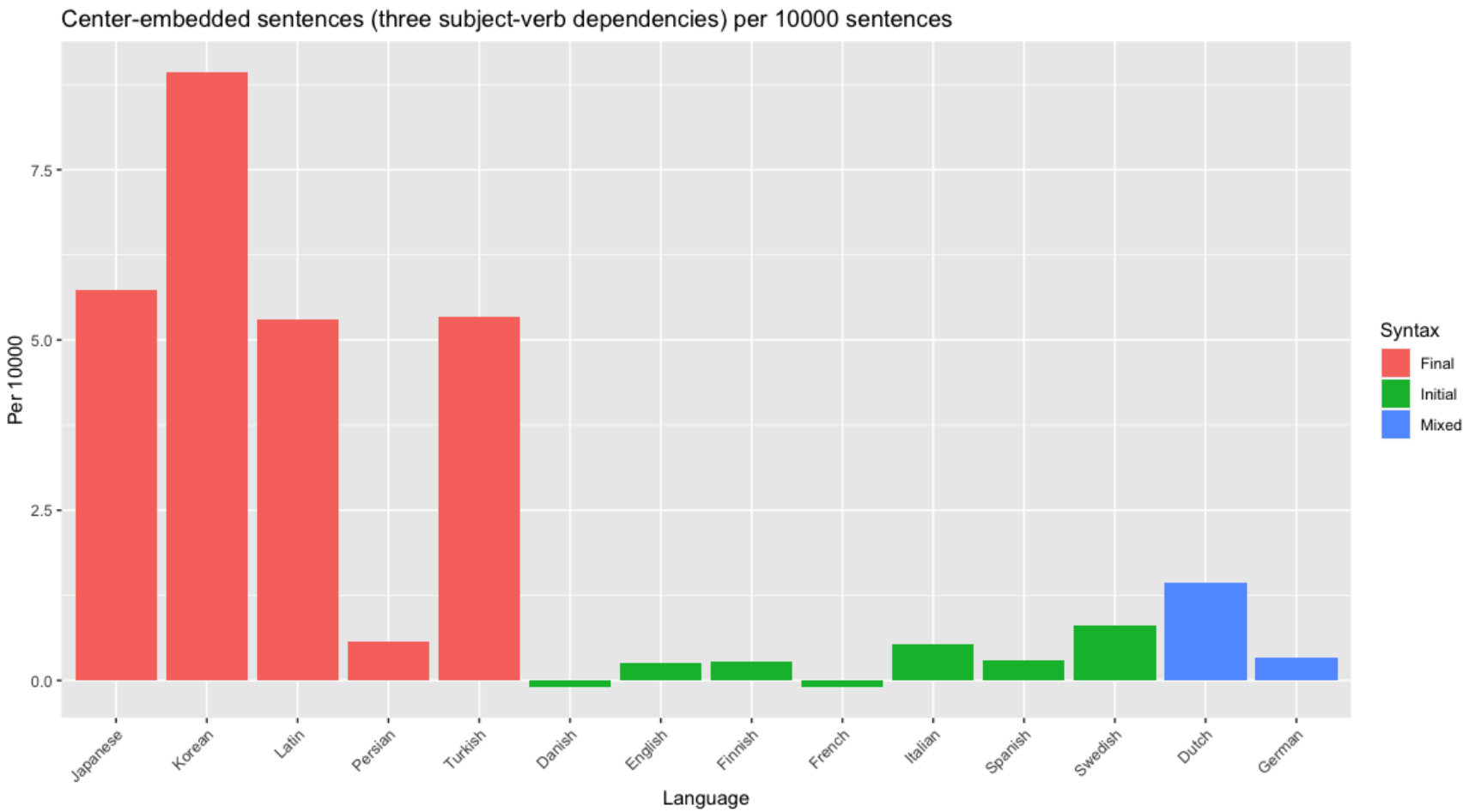

Figure 1. Two-level center-embedded sentences (per 10000) by language and syntax type