The information curve in verb-initial languages: a parallel corpus-based study

Since Greenberg (1963)'s seminal work on word order, it has been known that languages tend to disfavour verb-initial position (VSO and VOS) for their basic sentence and clause structure; in the most thorough survey on the order of Subject, Object, and Verb (1,376 languages), Dryer (2013) finds only 120 languages with verb-initial order.

From the seventies onward, typologists have suggested treating VSO and VOS languages together with SVO languages as they both have Verb-Object (VO) order, disregarding the position of the subject. Typologists have found that verb-initial languages differ little from verb medial languages in their structural features (Vennemann 1974; Vennemann 1976; Lehmann 1973; Lehmann 1978; Dryer 1991; Dryer 1992; Dryer 1997); as such, processing factors (Hawkins 2004) and/or grammaticalization patterns (Aristar 1991; Dryer 2019) that account for VO order would explain the VSO and VOS orders. Additionally, recent typological studies have questioned the verbal nature of the predicate in verb-initial languages, framing predicate fronting in terms of the more general problem of lexical flexibility (Gong and Uehara 2024).

Despite these functional or historical explanations, the fact remains that around 10% of the world's languages systematically choose a word order that is pragmatically marked in the other languages. In the absence of specific markers or syntactic constructions, the sentence initial position is used by many languages as the topic position. Furthermore, according to Klafka & Jurovski (2021) and contrary to the hypothesis of a smooth distribution of information in the sentence (Uniform Information Density: Levy & Jaeger 2007), the initial position of the sentence is characterized by a peak of information.

Klafka & Jurovski (2021) compute the curve of information in a sample of over 200 languages from Wikipedia, comparing sentences of the same length. We propose in the present study a comparison of the information curves of parallel sentences in a sample of languages from a corpus of literary texts, CIEP+. Whereas Klafka & Jurovski (2021) metrics are based on n-gram surprisal, we employ here surprisal estimates derived from mGPT (Shliazhko et al. 2024). For each of the commonly used predicate positions (initial, medial, final), we select three languages of different families: V-initial: Arabic, Irish, Tagalog; V-medial: Greek, Chinese, Finnish; V-final: Turkish, Japanese, Armenian. Working with a parallel corpus allows us to offer a better comparison of the curve of information across languages, as we are analyzing sentences with the same meaning and function. We expect to observe in verb-initial language a different shape of the information curve and to find similarities between verb-medial and verb-final languages; in the former languages, the peak of information is likely to be found in different positions and marked by specific constructions, while in the latter has a more fixed position, possibly, but not exclusive, at the beginning of the sentence.

References

Aristar, Anthony R. (1991). “On diachronic sources and synchronic pattern: An investigation into the origin of linguistic universals”. In: Language 67.1, pp. 1–33.

Dryer, Matthew S. (1991). “SVO Languages and the OV:VO Typology”. In: Journal of Linguistics 27, pp. 443–482. doi: 10.1017/S0022226700012743.

— (1992). “The Greenbergian Word Order Correlations”. In: Language 68, pp. 81–138. Doi: 10.2307/416370.

— (1997). “On the Six-Way Word Order Typology”. In: Studies in Language 21, pp. 69–103.

— (2013). “Order of Subject, Object and Verb (v2020.3)”. In: The World Atlas of Language Structures Online. Ed. by Matthew S. Dryer and Martin Haspelmath. Zenodo. doi: 10.5281/zenodo.7385533. url: https://doi.org/10.5281/zenodo.7385533.

— (2019). “Grammaticalization accounts of word order correlations”. In: Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence. Berlin: Language Science Press, pp. 63–95.

Gong, Liwei and Satoshi Uehara (2024). “Encoding of nominal predication constructions: a typological investigation in verb-initial languages”. In: Linguistic Typology 28 (2), pp. 291–330.

Greenberg, Joseph H. (1963). “Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements”. In: Universals of Human Language. Ed. by Joseph H. Greenberg. Cambridge, Mass: MIT Press, pp. 73–113.

Hawkins, J.A. (2004). Efficiency and Complexity in Grammars. Oxford linguistics. OUP Oxford.

Klafka, Josef, and Daniel Yurovsky. (2021). "Characterizing the Typical Information Curves of Diverse Languages" *Entropy* 23, no. 10: 1300. https://doi.org/10.3390/e23101300

Lascaratou, Chryssoula (1998). “Basic characteristics of Modern Greek word order”. In: Constituent Order in the Language of Europe. Ed. by Anna Siewierska. Berlin: Mouton de Gruyter, pp. 151–171.

Lehmann, Winfred P. (1973). “A structural principle of language and its implications”. In: Language 49, pp. 42–66.

— (1978). “The great underlying ground-plans”. In: Syntactic typology. Ed. by Winfred P. Lehmann. Austin: University of Texas Press, pp. 3–55.

Levy, R. and Jaeger, T.F. (2007) Speakers optimize information density through syntactic reduction. Adv. Neural Inf. Process. Syst., 19, 849.

Liu, Lei and Min Zhu (2022). “Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts”. In: Digital Scholarship in the Humanities 38.2, pp. 621–634. issn: 2055-7671. doi: 10.1093/llc/fqac089.

Payne, Doris L. (1990). The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages. Berlin: Mouton de Gruyter.

Shliazhko, Oleh et al. (2024). “mGPT: Few-Shot Learners Go Multilingual”. In: Transactions of the Association for Computational Linguistics 12, pp. 58–79. doi: 10.1162/tacl_a_00633.
url: https://aclanthology.org/2024.tacl-1.4.

Vennemann, Theo (1974). “Analogy in generative grammar: the origin of word order”. In: Proceedings of the Eleventh International Congress of Linguists (1972). Bologna: Il Mulino, pp. 79–83.

— (1976). “Categorial grammar and the order of meaningful elements”. In: Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday. Ed. by Alphonse Juilland. Saratoga, California: Anma Libri, pp. 615–634.