A Data-driven Approach to Crosslinguistic Structural Biases

Alex Kramer
University of Michigan, Ann Arbor
arkram@umich.edu

Zoey Liu
Boston College
ying.liu.5@bc.edu

Introduction. Ueno and Polinsky (2009) propose two structural biases that may facilitate processing efficiency: a pro-drop bias, which states that both SOV and SVO languages will use more pro-drop with transitive structures than with intransitive structures, and an intransitive bias, which states that SOV languages will use more intransitive structures than SVO languages. Corpus data comparing English and Spanish (SVO) to Japanese and Turkish (SOV) supported their predictions. Here, we expand upon these results by using naturalistic corpora and computational tools to investigate whether and to what extent subject drop (as opposed to pro-drop; see below) and intransitive biases are present at a larger cross-linguistic scale.

Hypotheses and predictions. Our hypotheses differ slightly from those of Ueno and Polinsky (2009) due to a key difference in method. We use a data-driven approach to determine presence of subject drop and transitivity: if a verb appears in an OV, VO, or V structure, the subject has been dropped (regardless of the particular grammatical or discourse reasons), and if a verb appears in an SV or VS structure, it is intransitive. In contrast, in Ueno and Polinsky (2009), the transitivity of each verb was annotated manually. This is a potential issue because there is not a clear cross-linguistic distinction between object drop and intransitivity, indicating in turn that the presence or absence of object drop in their study was decided in a more subjective manner.

Family	# of languages
Afro-Asiatic	4
Dravidian	2
Indo-European (IE)	40
Niger-Congo	2
Sino-Tibetan	2
Turkic	2
Uralic	5
Other	9

Table 1: Breakdown of languages in dataset by family.

An additional caveat of the method employed by Ueno and Polinsky (2009) is their treatment of word order. In this study, word order was coded categorically as either SOV or SVO. However, the use of categorical typological variables can lead to data reduction and, consequently, statistical bias, for example in the form of bimodal distributions (Wälchli, 2009). Computational analysis using gradient measures of word order can reduce bias and allow for testing more fine-grained predictions (Levshina, 2019). We thus examine how both categorical (dominant word order) and continuous measures (headedness) predict these two biases.

Given our method, we hypothesized that languages traditionally categorized as predominantly or rigidly SOV (e.g. Hindi) would show a stronger bias toward subject drop and make greater use of intransitive structures than languages categorized as SVO. Similarly, we hypothesized that languages that are not necessarily categorized as SOV but nonetheless have a high degree of head finality (e.g. Mandarin[1]), would show a stronger bias toward subject drop and make greater use of intransitive structures than more head-initial languages.

With a categorical approach, the subject drop bias predicts that SOV languages will have a higher proportion of OV/VO/V structures than SVO languages, and the intransitive bias predicts that SOV languages will have a higher proportion of SV/VS structures than SVO languages. With a gradient approach, the subject drop bias predicts that there will be a positive correlation between proportions of OV/VO/V structures and proportions of head-final dependencies, and the intransitive bias predicts that there will be a positive correlation between proportions of SV/VS structures and proportions of head-final dependencies.

Data and preprocessing. We used treebanks of

[1] Mandarin is categorized as SVO but has a head-finality proportion of 0.63, similar to SOV languages such as Turkish (0.66) and Telugu (0.61) (see **Data and preprocessing**).

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 344-346.
Held on-line February 14-19, 2021

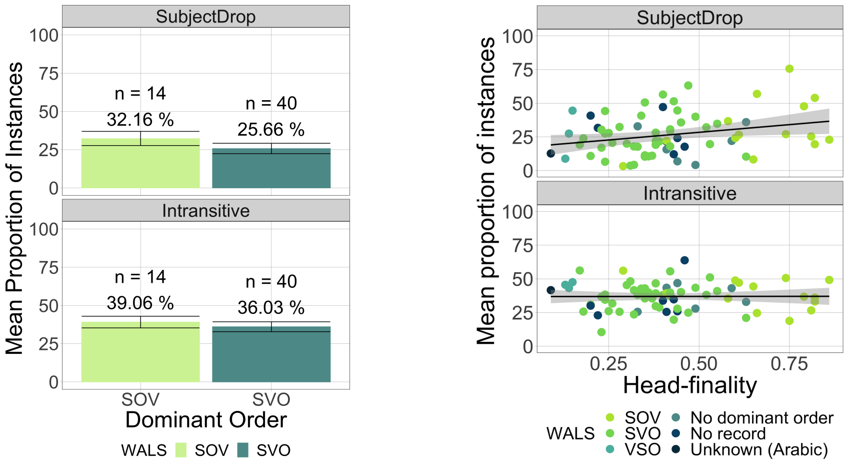

Figure 1: *Left*: Mean proportion of instances of subject drop and intransitive sentences in languages categorized as SOV and SVO. *Right*: Mean proportion of instances of subject drop and intransitive sentences by proportion of head-final dependencies. Arabic is coded as "Unknown" because the variety represented in UD 2.6 was unclear.

contemporary languages from the Universal Dependencies version 2.6 (Zeman et al., 2020). We extracted all sentences without S or O (SV, VS, OV, VO, V) where V was the root of the sentence. Languages with under 100 combined instances of SV, VS, OV, VO, and V structures were excluded, leaving a dataset of 66 languages (Table 1). Certain languages, such as Norwegian, included multiple varieties, which were treated separately.

The proportion of each order to the total number of sentences was calculated, and head-finality was calculated as the proportion of head-final dependencies to all other dependencies, excluding function words (Futrell et al., 2020). Significance testing was performed via bootstrapping (Efron, 1979) for 10,000 iterations. The dominant order of each language was additionally coded via WALS (Dryer and Haspelmath, 2013) for comparison.

Results. There was no significant difference between proportions of subject drop in SOV and SVO languages (μ_{sov} = 32.16, CI = [27.70, 36.95]; μ_{svo} = 25.66, CI = [22.36, 29.16]). However, there was a significant positive correlation between subject drop and head-finality (ρ = 0.27, p = 0.02, CI = [0.04, 0.47]). In other words, languages in our sample with more head-final dependencies tended to contain more instances of subject drop than those with fewer head-final dependencies, regardless of their traditionally-classified dominant orders.

There was also no significant difference between proportions of intransitive structures in SOV and SVO languages (μ_{sov} = 39.06, CI = [35.32, 42.95]; μ_{svo} = 36.03, CI = [32.83, 39.32]), nor was there a significant correlation between intransitive structures and head-finality (ρ = 0, p = 0.97, CI = [-0.23, 0.24]). In other words, intransitive structures were not more common among the SOV languages in our sample, nor were they more common among more head-final languages. This calls into question the hypothesis proposed by Ueno and Polinsky (2009) that intransitive structures are exploited by SOV languages to facilitate processing.

Conclusion. By traditional categorization, SVO and SOV languages in our data set did not differ significantly in their use of subject drop or intransitives. Only when we considered head-finality as a gradient measure did a trend emerge, with more head-final languages using more subject drop.

With a much smaller data set, Ueno and Polinsky (2009) demonstrated that both SVO and SOV contexts showed a pro-drop bias, but only SOV contexts presented an intransitive bias. However, that was not the case at least with the structures that we have studied. Controlled online experiments should be carried out to further test the relationship between these structures and processing efficiency.

Our ongoing work focuses on extending the current study through parallel corpora[2] in order to determine whether these patterns hold when meaning is controlled for. In addition, the present data is heavily skewed toward IE languages, potentially leading to issues of non-independence. Parallel corpora will allow us to incorporate data from a wider variety of languages, reducing this skew.

[2] http://www.statmt.org/wmt20/translation-task.html

References

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bradley Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96:371–412.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Mieko Ueno and Maria Polinsky. 2009. Does headedness affect processing? A new look at the VO–OV contrast. *Journal of Linguistics*, 45(3):675–710.

Bernhard Wälchli. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology*, 13(1).

Daniel Zeman, Joakim Nivre, and Mitchell Abrams et al. 2020. Universal dependencies 2.6. LINDAT/ CLARIAH-CZ, Faculty of Mathematics and Physics, Charles University.