Defining constituent order flexibility from a typological perspective: WALS, AUTOTYP, and beyond

How does constituent order vary cross-linguistically, and what drives this variation? Large-scale typological databases such as WALS (Dryer & Haspelmath 2013) and AUTOTYP (Bickel et al. 2017) have focused on cataloging the dominant constituent orders of the world's languages. However, languages vary not only in their primary order(s), but also in the number of additional orders speakers accept and the degree to which they accept them—their flexibility (Namboodiripad 2017). Here, we compare the criteria used by each database in determining (non)dominant constituent order and argue that expanding existing notions of flexibility can lead to important insights about this variation and its sources.

Differences in how the databases determine category membership illustrate the challenges associated with categorical approaches to constituent order. WALS uses corpus data to determine DOMINANT WORD ORDER, which is defined as the order which occurs at least twice as often as the next most frequent order. If no corpus exists, a grammar is consulted instead. AUTOTYP, using grammars, additionally classifies languages as RIGID, FLEXIBLE, or FREE: rigid languages have exactly one basic order, flexible languages have a basic order and one or more structurally-conditioned orders, and free languages have no basic order. There was significant overlap in the classifications in these databases (N=266; 85%). Of the 46 non-overlapping languages, 28 (61%) constituted true disagreements, 8 (17%) were classified differently from each other due to differing definitions, and 10 (22%) were unclear due to the use of different language varieties.

These differences notwithstanding, the information in such databases can point us toward potential correlates of flexibility. We aggregated the constituent order data in WALS and AUTOTYP alongside a set of additional features we predicted would pattern with flexibility: grammatical case-marking, argument marking on the verb, the use of head- or dependent-marking, and the presence of pro-drop. In line with previous work, we found flexible languages to be somewhat more likely to have case marking, and rigid languages more likely to lack argument marking on the verb (Figures 1 and 2).

However, manual inspection of AUTOTYP's "rigid" category gave us pause: There is a sense in which many of these languages are not strictly rigid. For example, Russian is classified as a rigid SVO language, yet intuitively, it does not pattern with English, another rigid SVO language; all six orderings of major constituents are grammatical and attested in Russian (Bailyn 2012), while this is not the case in English. Likewise, many SOV languages—for example, Korean—which allow all of the logical constituent orders are classified as rigid SOV, even though they exhibit considerable (discourse-mediated) flexibility, as shown in experimental work (Namboodiripad, Kim, & Kim 2019).

We conclude with a comparison of three languages classified as SOV flexible (Avar, Korean, and Malayalam) which nonetheless exhibit subtle differences in flexibility. We propose that supplementing existing discrete categories such as "flexible" and "rigid" with a gradient notion of flexibility increases descriptive power and, with enough data, could improve correlational investigations of constituent order typology.

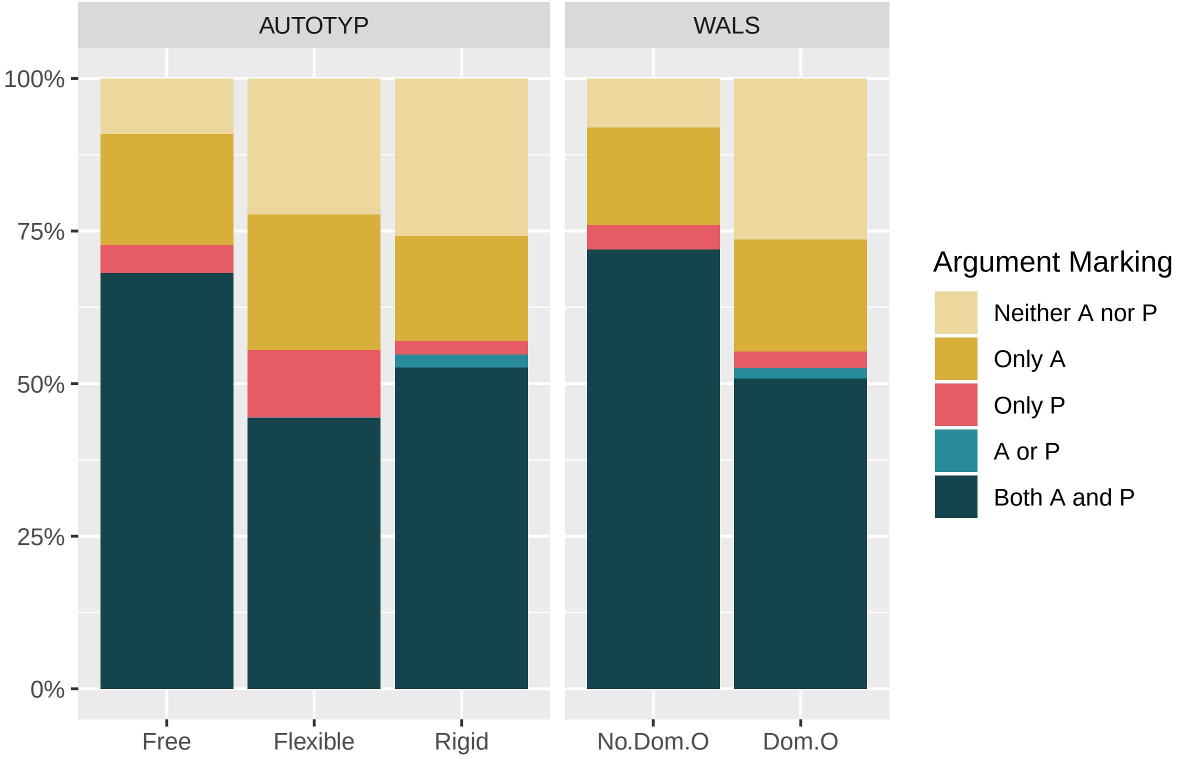

Figure 1: Percentage of languages with different types of argument marking

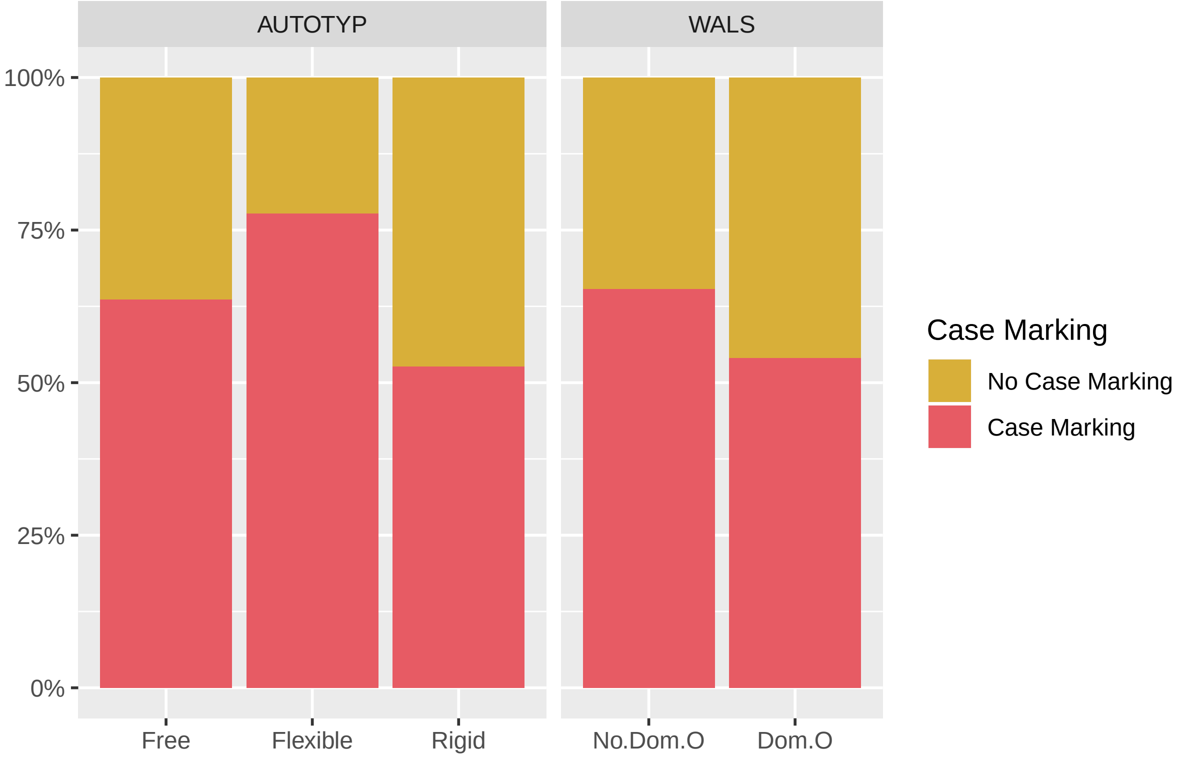

Figure 2: Percentage of languages with and without case marking