THE ORIGINS OF WORD ORDER UNIVERSALS: EVIDENCE FROM CORPUS STATISTICS AND SILENT GESTURE

SIMON KIRBY, JENNIFER CULBERTSON, AND MARIEKE SCHOUWSTRA

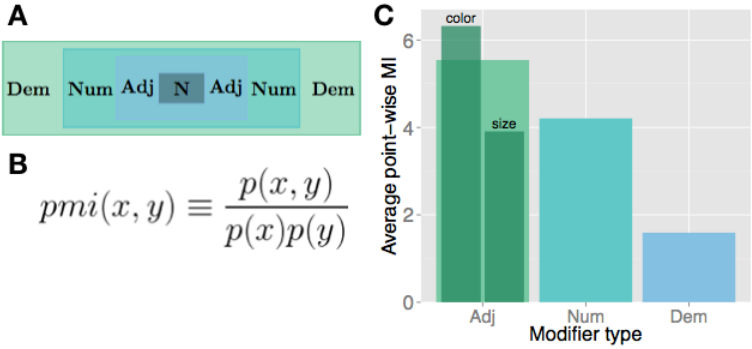
Centre for Language Evolution, University of Edinburgh, UK
simon@ling.ed.ac.uk, jennifer.culbertson@ed.ac.uk, marieke.schouwstra@ed.ac.uk

Why are some patterns of noun phrase order (e.g., N-Adj-Num-Dem, 'houses big five these') much more common than others (e.g., N-Dem-Num-Adj, 'houses these five big')? An intriguing possibility is that this distribution emerges in response to a general cognitive bias favouring a transparent relationship between conceptual structure and linear order – an isomorphism bias. In conceptual structure, Adj is closest to N, then Num, then Dem (Fig. 1A). Linear orders that can be read off this nested structure are isomorphic, and these orders are the most common cross-linguistically. Previous experimental work has found an isomorphism in both traditional artificial language learning (Culbertson & Adger, 2014), and silent gesture experiments (Culbertson et al., 2016). For example, Culbertson et al. (2016) asked non-signing English speakers to communicate simple pictures using only gesture. Items were groups of 4 or 5 (Num) objects (N), either spotted or striped (Adj), in a proximal or distal location (Dem). Participants spontaneously improvised isomorphic gestures that did not reflect their native language. These results suggest that an isomorphism bias is present, however they leave open the *origin* of this bias. In particular, they cannot tell us the origins of the conceptual structure that word order is isomorphic to. Here we show that the conceptual structure of the noun phrase is learnable by observing simple statistics about objects in the world.

Intuitively, our proposal is this: properties (~Adj) are more inherent to objects than numerosities (~Num), and location or discourse status (~Dem) is generally not an inherent feature of objects (cf. Rijkhoff 2002). More precisely, we quantify this notion using *point-wise mutual information* (Fig. 1B), a measure of the strength of association between pairs of elements. Using linguistic corpora as a proxy for the world, we can measure average pmi

DOI:10.12775/3991-1.051

between nouns and each class of modifier (within a certain window). By averaging over all pairs of, for example, numerals and nouns, we can get an overall measure of how inherent numerosity is to objects. Large corpora of 8 different languages (English, German, French, Spanish, Italian, Portuguese, Chinese, Arabic), plus all English child-directed corpora in CHILDES, confirm our intuition: average pmi of Adj and Num were higher than Dem in all cases, and average pmi of Adj was higher than Num in all but the Portuguese corpus (e.g., Fig. 1C). The conceptual structure in Fig. 1A is thus likely learnable from properties of the world. Interestingly, our measure of inherentness derives a new prediction about sub-classes of adjectives (cf. Martin 1969, Bouchard 2002). Specifically, the corpus results suggest that *size* may not be as inherent to objects as *texture/color* (which pattern together, see Fig. 1C). This predicts that size adjectives may show a weaker isomorphism bias.



$$pmi(x, y) \equiv \frac{p(x, y)}{p(x)p(y)}$$

Fig 1. **A:** conceptual structure. **B:** point-wise mutual information. **C:** average pmi across modifier types, and adjective subclasses in English (Brown corpus).

To test this, we adapt Culbertson et al.'s (2016) silent gesture experiment, swapping *big* or *small* for the original adjectives. We replicate the original findings for isomorphic order of Num and Dem, but as predicted, the preference for isomorphic order of Adj relative to Num (and to some extent even Dem) is weakened. In other words, isomorphism can be modulated by average pmi, a measure of relative inherentness. This supports the claim that the isomorphism bias taps into a conceptual structure reflecting statistical properties of the world.

In summary, the underlying conceptual structure of the noun phrase—which shapes the typological distribution of orders in this domain—is learnable from observing the statistical properties of the world: it reflects different strengths of associations between objects and their properties, numerosities and locations.

References

Bouchard, D. (2002). Adjectives, number and interfaces: Why languages vary. Elsevier, Amsterdam.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16), 5824-5847.

Culbertson, J., Kirby, S., & Schouwstra, M. (2016). Word order universals reflect cognitive biases: Evidence from silent gesture. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barcelo-Coblijn, O. Feher & T. Verhoef (eds.) *The Evolution of Language: proceedings of the 11th International Conference*. doi:10.17617/2.2248195

Martin, J. E. (1969). Semantic determinants of preferred adjective order. Journal of Verbal Learning and Verbal Behavior, 8(6):697–704.

Rijkhoff, J. (2002). The noun phrase. Oxford University Press, USA.