Natalia Levshina/Adèle H. Ribeiro

"Who did what to whom":

Measuring and explaining cross-linguistic differences

Keywords: case marking; word order; causal analysis

Different languages use different linguistic cues to express "who did what to whom", helping the addressee to identify Subject and Object. These cues include case marking, agreement, semantics, and word order. Previous research has revealed that different cues can be correlated (Greenberg 1966; Sinnemäki 2010; Levshina 2021). For example, some languages express the roles with case (Latin, Czech) and relatively flexible word order, while others (English, Mandarin) use rigid word order and have no nominal case makers. The differences between the languages have been explained by sociolinguistic factors, such as population size and high proportion of L2 (non-native) users, which can lead to grammatical simplification – in particular, to loss of case (Lupyan/Dale 2010; Trudgill 2011; Bentz & Winter 2013; Koplenig 2019) and increased use of verb-medial order (Lev-Ari 2023). In this paper, we measure the differences between languages with the help of typological databases and corpus data, and explain these differences by learning potential causal relationships among linguistic and sociolinguistic variables with the help of cutting-edge causal inference techniques (Pearl 2000).

We used the World Atlas of Language Structures (WALS) (Dryer/Haspelmath 2013), the parallel corpus of Bible translations (Mayer/Cysouw 2014) and word order data inferred from this corpus (Östling 2015). From these sources we obtained information about three variables: 1) "SO Entropy", the entropy of Subject and Object order based on the probabilities of Subject-Object and Object-Subject orders in the corpus; 2) "Case", whether case flagging helps to distinguish between the forms of Subject and Object; and 3) "Verb Final", whether the position of the lexical verb is final or non-final. We also used information about the population size from Koplenig (2019). Overall, we obtained linguistic and sociolinguistic data for 827 languages representing 78 language families.

Next, we performed correlational and causal analyses between the variables. To discover potential causal (ancestral) relationships among these variables, we applied the Fast Causal Inference (FCI) algorithm (Zhang 2008). FCI learns from data a Partial Ancestral Graph (PAG) that represents the class of all causal models, potentially involving unobserved confounders, that explain the observed conditional independencies, referred to as Markov Equivalence Class (MEC). Ancestral and non-ancestral relationships that are shared among all models in the MEC are represented in the PAG by non-circle edge marks (i.e., tails and arrowheads, respectively). Since our dataset includes variables of mixed types (numeric and categorical), we constructed a conditional independence test based on fitting mixed-effects regression models. The genealogical and geographic dependencies between the languages were controlled by treating the genera and macroareas as random intercepts.

A PAG representing plausible (non-)ancestral relationships between the variables is shown in Figure 1. The model indicates that population size is associated with Case and Verb Final, which is consistent with previous studies. Moreover, it suggests that Case and Verb Final are not ancestors or underlying causes of SO Entropy or Population Size, which also aligns with prior research findings.

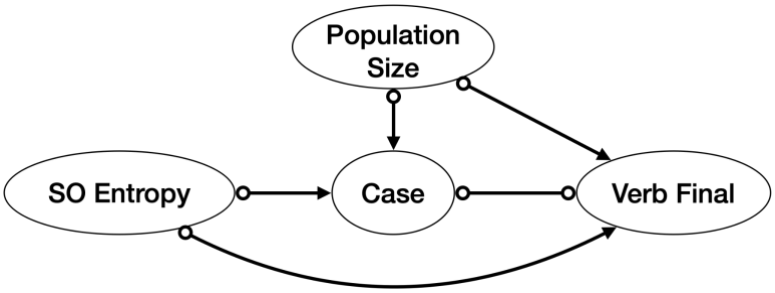

Fig. 1: A Partial Ancestral Graph (PAG) resulting from our causal analysis describing the (non-)ancestral relationships among linguistic variables and population size.

References

Bentz, Christian/Winter, Bodo (2013): Languages with more second language learners tend to lose nominal case. In: Language Dynamics and Change, 3, pp. 1–27.

Dryer, Matthew/Haspelmath, Martin (2013): The world atlas of language structures online. http://wals.info (last access: 10 May 2023).

Greenberg, Joseph (1966): Some universals of grammar with particular reference to the order of meaningful elements. In: J. Greenberg (ed.), Universals of Grammar, Cambridge, MA: MIT Press, pp. 73–113.

Koplenig, Alexander (2019): Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. In: Royal Society Open Science, 6, 181274.

Lev-Ari, Shiri (2023): The emergence of word order from a social network perspective. In: Cognition, 237, 105466.

Levshina, Natalia (2021): Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. In: Frontiers in Psychology, 12, 648200.

Lupyan, Gary/Dale, Rick (2010): Language structure is partly determined by social structure. In: PLoS One, 5, e8559.

Mayer, Thomas/Cysouw, Michael (2014): Creating a massively parallel bible corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, pp. 3158–3163.

Östling, Robert (2015): Word order typology through multilingual word alignment. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, pp. 205–211.

Pearl, Judea (2000): Causality: Models, Reasoning, and Inference. New York: Cambridge University Press.

Sinnemäki, Kaius (2010): Word order in zero-marking languages. In: Studies in Language, 34, pp. 869–912.

Trudgill, Peter (2011): Sociolinguistic typology: Social determinants of linguistic complexity. Oxford: Oxford University Press.

Zhang, Jiji (2008): On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. In: Artificial Intelligence, 172(16-17), pp. 1873–1896.

Contact information

Natalia Levshina

Max Planck Institute for Psycholinguistics

natalevs@gmail.com

Adèle H. Ribeiro

Philipps-Universität Marburg

adele.ribeiro@uni-marburg.de