"Who did what to whom": Measuring and explaining cross-linguistic differences

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; McWhorter 2011; Trudgill 2011; Bentz & Winter 2013; Koplenig 2019). In this paper, we measure the differences between languages with the help of typological databases and corpus data. Our goal is also to 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 diverse sources of data: large web-based corpora of online news (Goldhahn et al. 2012) annotated with Universal Dependencies; Universal Dependencies corpora (Zeman et al. 2022); the parallel corpus of Bible translations (Mayer & Cysouw 2014) and word order data inferred from this corpus (Östling 2015); and the World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2013). From these sources we obtained information about three variables that help to understand "who did what to whom", according to the literature: 1) the entropy of Subject and Object order based on the probabilities of Subject-Object (SO) and Object-Subject (OS) orders in the corpora; 2) whether the forms of Subject and Object are the same or distinct thanks to case flagging; and 3) the position of the lexical verb in a transitive clause: final or non-final. We also used information about the population size and L2 speaker proportions from the Ethnologue database, as well as the datasets from Koplenig (2019) and Sinnemäki & Di Garbo (2018). Overall, we have managed to obtain the linguistic and sociolinguistic data for 112 languages representing 45 genera.

Next, we compared the languages and performed correlational and causal analyses between six variables, namely *Macroarea* (the geographical area where the language emerged), *Total Users* (the total number of language users), *L2 Prop* (the proportion of non-native language speakers), *SO_Form* (the form in which subject and object appear in the sentence), *SO_Entropy* (the entropy of the Subject-Object order), and *Verb* (the position of the verb in the sentence in relation to Subject and Object). To discover potential causal (ancestral) relationships among these variables, we applied the Fast Causal Inference (FCI) algorithm (Zhang 2008) to the synchronic data. The FCI algorithm recovers a Partial Ancestral Graph (PAG) that represents the class of all possible causal models that explain the conditional independencies observed in the data, named Makov Equivalence Class, accounting for the presence of unmeasured confounders. Since our dataset includes variables of mixed types (numeric and categorical), we constructed a conditional independence test based on fitting mixed-effects regression models (namely, Gaussian, logistic binomial, and beta regression). The genealogical dependencies between the languages were controlled by treating the genera as random intercepts. The macroareas were treated as fixed effects.

Our first results demonstrate that the linguistic cues are moderately correlated, supporting the previous studies. A PAG obtained by our preliminary results is displayed in Figure 1. The model suggests that *Area* and *Verb* are ancestors (causes) of *SO Form*, which can be interpreted theoretically. It further suggests that Area is not an ancestor (and, therefore, not a cause) of *L2 Prop* and *Total Users*. Possibly, Area is associated with *L2 Prop* and *Total Users* due to only unmeasured confounders. We also observe some expected associations between *Total Users*, *SO_Entropy* and *Verb*.

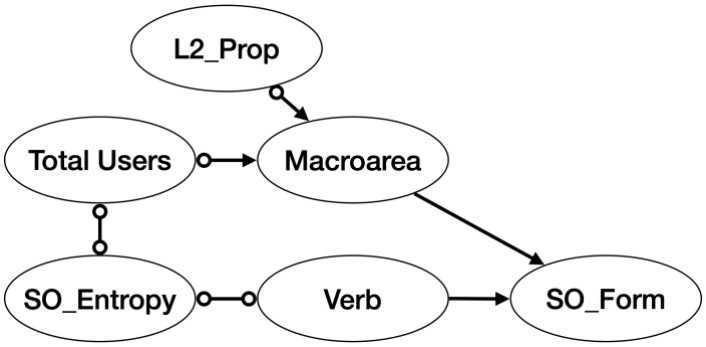

Figure 1. A Partial Ancestral Graph (PAG) resulting from our preliminary causal analysis describing the potential (non-)ancestral relationships among linguistic variables.

References

Bentz, Ch. & B. Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3: 1–27.

Dryer, M. & M. Haspelmath. 2013. The world atlas of language structures online. http://wals.info.

Goldhahn, D., Th. Eckart & U. Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 759–765, Istanbul, 2012.

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

Hawkins, J. A. 1986. *Comparative Typology of English and German. Unifying the contrasts*. Croom Helm, London, 1986.

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

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

Lupyan, G. & R. Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5:e8559.

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

McWhorter, J. 2011. Linguistic simplicity and complexity: Why do languages undress? Berlin: de Gruyter Mouton.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York.

Robert Östling. 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), pp. 205–211. Beijing.

Sinnemäki, K. 2010. Word order in zero-marking languages. *Studies in Language* 34: 869–912.

Sinnemäki, K. & F. Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9: 1141.

Trudgill, P. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Zeman, D. et al. 2022. Universal dependencies 2.10. http://hdl.handle.net/11234/1-4758.

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