A cautionary note on sociodemographic predictors of linguistic complexity: different measures and different analyses lead to different conclusions

Gary Lupyan[*1] and Limor Raviv[2,3]

[*]Corresponding Author: lupyan@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
[2]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[3]cSCAN, University of Glasgow, Glasgow, UK

The question of why languages differ in the ways they do has been of long-standing interest in the fields of language evolution and language diversity. In 2010, Lupyan and Dale took advantage of the recently digitized World Atlas of Language Structures (WALS; Haspelmath et al., 2008) to test a hypothesis variously articulated by Trudgill (2002), Wray and Grace (2007) and McWhorter (2007). The broader claim was that some linguistic differences may arise from languages "adapting" to different sociodemographic environments. The more specific claim was that languages with histories of use by larger and more diverse speaking populations will tend to lack features such as complex agreement and inflectional systems that are thought to be difficult for nonnative learners (i.e., outsiders) to master. In line with this hypothesis, Lupyan & Dale, 2010 found that languages spoken by more people and spread over a larger area were more likely to use more lexical rather than inflectional means of communicating various information such as aspect, evidentiality, and possibility, and to systematically have fewer grammatical distinctions in e.g., types of possession, remoteness distinctions in tense, and grammatical encoding of space in demonstratives.

Numerous correlational studies have since confirmed this general trend (e.g., Bentz & Winter, 2013; Bentz et al., 2015; Nettle, 2012), but new analyses come with new caveats, e.g, some finding that the proportion of L2 speakers matters (Sinnemäki & Garbo, 2018), while others finding that it does not (Koplenig, 2019). Importantly, the link between group size and language structures has also begun to be experimentally tested, with studies finding that larger groups produce more systematic and compositional structure (Raviv et al., 2019, 2020). At the same time, developmental studies have been finding some evidence that children learn better from more redundant (i.e., complex) language input (Tal & Arnon, 2022; Portelance et al., 2023), helping to explain why languages may end up with such high levels of redundancy in the first place.

Recently, Shcherbakova et al. (2023) conducted a meticulous analysis of the link between grammatical complexity and sociodemographic factors and came

to a very different conclusion, finding either no link or a *positive* relationship between complexity and population size. Their conclusion–"societies of strangers do not speak less complex languages"–squarely contradicts the earlier results.

We conducted a reanalysis to better understand what accounted for the qualitative difference between earlier work and Shcherbakova et al.'s (2023) results. Compared to past work, Shcherbakova et al. used more complex areal and phylogenetic models to better control for non-independence of languages. But what is so puzzling is that their analysis failed to find a negative association between population size and complexity even in the raw data, *prior* to the areal and phylogenetic controls suggesting that the difference in phylogenetic controls was the main source of the discrepancy. Neither were the sociodemographic predictors since these matched earlier work. This leaves two main sources of difference: which languages were included in the sample, and how grammatical differences/complexity was quantified.

Shcherbakova et al.'s analysis used the newly available Grambank database (Skirgård et al., 2023). Despite including 1314 languages, the new sample only partially overlaps with the earlier WALS-based sample. For example, Shcherbakova et al's data includes long-extinct languages such as Ancient Hebrew and Ancient Greek (with 0 speakers) while omitting some large languages such as German, Spanish, Bengali, and Gujarati (which are included in WALS). Because very small-population and very large-population languages have larger influence on the regression models predicting complexity from (log-transformed) population size, even small differences in which languages are included can lead to large differences in results.

In WALS, grammatical features are coded using ordinal, nominal, and binary schemes while Grambank codes all features as binary. The binary coding simplifies analysis, but can radically change the relative weighing of variables, e.g., what was previously one variable (number of cases) becomes *n* variables (has ablative, has locative, etc.). In our re-analysis, we found that the features associated with smaller populations tended to be those previously described, e.g., more complex possessive markings (GB058, GB059), demonstratives (GB036), and distinctions in clusivity (GB028) and pronouns (GB0310). The features found to be associated *positively* with population such as politeness (GB415), diminutives (GB315), and passive markers (GB147) were not included in earlier analyses, helping to further explain the discrepancies in the results.

Different language databases, different measures, and different analyses can yield substantially different conclusions about the relationship between sociodemographic features and linguistic complexity. Our re-analysis suggests that Shcherbakova et al's (2023) does not supersede earlier results, but the broader coverage offered by Grambank combined with the more powerful areal and phylogenetic controls can be used to gain further insights into which aspects of language are and are not shaped by sociodemographic factors.

Acknowledgements

We thank Olena Shcherbakova for sharing the data and analyses with us.

References

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PloS One*, *10*(6), e0128254.

Bentz, C., & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, *3*(1), 1–27.

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (2008). *The world atlas of language structures online.* Munich: Max Planck Digital Library.

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*(2), 181274.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1), e8559.

Mcwhorter, J. (2007). *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars.* Oxford University Press, USA.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting Age of Acquisition for Children's Early Vocabulary in Five Languages Using Language Model Surprisal. *Cognitive Science*, *47*(9), e13334.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The Role of Social Network Structure in the Emergence of Linguistic Structure. *Cognitive Science*, *44*(8), e12876.

Shcherbakova, O., Michaelis, S. M., Haynie, H. J., Passmore, S., Gast, V., Gray, R. D., Greenhill, S. J., Blasi, D. E., & Skirgård, H. (2023). Societies of strangers do not speak less complex languages. *Science Advances*, *9*(33), eadf7704.

Sinnemäki, K., & Garbo, F. D. (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*, 342569.

Skirgård, H., et al.. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175.

Tal, S., & Arnon, I. (2022). Redundancy can benefit learning: Evidence from word order and case marking. *Cognition*, *224*, 105055.

Trudgill, P. (2002). *Sociolinguistic Variation and Change*. Georgetown University Press.

Wray, A., & Grace, G. (2007). The consequences of talking to strangers : Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.