INVESTIGATING THE RELATIONSHIP BETWEEN NUMBER OF SPEAKERS, I-COMPLEXITY, AND E-COMPLEXITY

Arturs Semenuks*[*1,2]

[*]Corresponding Author: asemenuks@ucsd.edu
[1]Department of Cognitive Science, UCSD, La Jolla, USA
[2]Center for Academic Research and Training in Anthropogeny, UCSD, La Jolla, USA

A number of theoretical proposals and computational models suggest that the sociocultural niche a language occupies affects the morphosyntactic complexity of that language, e.g. see McWhorter (2001), Wray and Grace (2007), Trudgill (2011), Dale and Lupyan (2012), Spike (2017). A common denominator of many such proposals is the focus on the difference between exoteric and esoteric societies, i.e. societies more (exoteric) and less (esoteric) open to outsiders. Wray and Grace (2007) propose arguably the most detailed psycholinguistic explanation for how esotericity promotes morphosyntactic complexity to date, suggesting that shared insider knowledge and implicit encoding leads to opaque, irregular, and more complex forms of language[1].

Recently, correlational cross-linguistic and experimental studies have provided empirical support to these theories, e.g. Lupyan and Dale (2010), Bentz and Winter (2014), Sinnemäki and Di Garbo (2018), Koplenig (2019), Kocab, Ziegler, and Snedeker (2019), Raviv, Meyer, and Lev-Ari (2019), Berdicevskis and Semenuks (2020, 2022). It is encouraging that researchers arrive at similar conclusions using approaches differing in methodological details, which could be taken as showing the robustness of the hypothesized relationship. At the same time, what morphological complexity is and what measures best capture it is not yet settled (Berdicevskis et al., 2018; Ehret, Blumenthal-Dramé, Bentz, & Berdicevskis, 2021). Additionally, how different dimensions of complexity interact with each other is far from resolved and what specific dimension(s) of complexity are affected by sociocultural niche is not yet clear. For example, Sinnemäki and Di Garbo (2018) find their verbal morphological complexity measure to be correlated with the total amount of speakers a language has and the percentage of L2 speakers in its population, but find no similar relationship for nominal complexity. Thus the aforementioned variability in methodological details leaves open

[1]In turn, these properties themselves make it harder for outsiders to understand the language, creating a feedback loop.

some important questions, including (i) are languages in exoteric niches simpler on all dimensions of complexity? and (ii) how are different dimensions of complexity related?

Here I explore whether languages of more exoteric societies (operationalized as the total number of speakers) tend to have morphological paradigms of lower i-complexity. Following Ackerman and Malouf (2013), we define i-complexity as the average conditional entropy between the word forms in a paradigm, which captures the average amount of information needed for a speaker to predict all inflectional forms of a new lemma. On the one hand, it should be expected that lower values of i-complexity facilitate learning and would be under stronger selective pressure in exoteric societies with more speakers. On the other hand, some information theoretic measures of language structure have been reported to be highly similar across languages due to psycholinguistic constraints (Coupé, Oh, Dediu, & Pellegrino, 2019). As Ackerman and Malouf (2013) hypothesize that i-complexity is constrained to facilitate preservation of efficient linguistic structures, we can similarly expect it to not differ substantially across languages.

Additionally, recent studies have provided evidence for a negative correlation between i-complexity and e-complexity, i.e. the number of morphosyntactic distinctions that a particular language makes (Cotterell, Kirov, Hulden, & Eisner, 2019; Johnson, Gao, Smith, Rabagliati, & Culbertson, 2021). I extend the previous analyses using methods more closely aligned with the proposal in Ackerman and Malouf (2013) and following the advice for best practices and directions in Malouf, Ackerman, and Semenuks (2020).

I use the data (71 languages, 47 language families) from the Surrey Morphological Complexity Database (Baerman, Brown, Evans, Corbett, & Cahill, 2015), as well as the data from the UniMorph 3.0 Project (60 languages, 11 language families) (McCarthy et al., 2020). The former database provides high quality annotated data for inflectional paradigms from a typologically and geographically diverse sample of languages, whereas the latter provides a less diverse sample, but allows for a more theory-neutral data-driven process of estimating paradigm information for a language. In the UniMorph 3.0 data, the set of noun paradigms for each language is estimated by removing the maximal shared subset of characters within word forms for all lemmas in a language. I use mixed-effect linear models to control for phylogeny and geography in analyzing the data. No significant correlation between i-complexity and the number of speakers a language is observed. However, similarly to Ackerman and Malouf (2013), languages are found to frequently have lower values than expected as shown through Monte Carlo simulations. Additionally, both data sets suggest a negative correlation between i-compexity and e-complexity. Taken together, the results suggest that i-complexity is not significantly affected by the sociocultural niche a language occupies, however we see evidence of optimization of this aspect of linguistic structure, potentially supporting its hypothesized psycholinguistic importance.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 429–464.

Baerman, M., Brown, D., Evans, R., Corbett, G. G., & Cahill, L. (2015). *Surrey morphological complexity database.* (University of Surrey. `http://dx.doi.org/10.15126/SMG.23/1`)

Bentz, C., & Winter, B. (2014). Languages with more second language learners tend to lose nominal case. In *Quantifying language dynamics* (pp. 96–124). Brill.

Berdicevskis, A., Çöltekin, Ç., Ehret, K., Prince, K. von, Ross, D., Thompson, B., Yan, C., Demberg, V., Lupyan, G., Rama, T., et al.. (2018). Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the second workshop on universal dependencies (udw 2018).*

Berdicevskis, A., & Semenuks, A. (2020). Different trajectories of morphological overspecification and irregularity under imperfect language learning. *The Complexities of Morphology*, 283–305.

Berdicevskis, A., & Semenuks, A. (2022). Imperfect language learning reduces morphological overspecification: Experimental evidence. *PloS one*, *17*(1), e0262876.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, *7*, 327–342.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, *5*(9), eaaw2594.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems*, *15*(03n04), 1150017.

Ehret, K., Blumenthal-Dramé, A., Bentz, C., & Berdicevskis, A. (2021). Meaning and measures: interpreting and evaluating complexity metrics. *Frontiers in Communication*, *6*, 61.

Johnson, T., Gao, K., Smith, K., Rabagliati, H., & Culbertson, J. (2021). Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling*, *9*(1), 97–150.

Kocab, A., Ziegler, J., & Snedeker, J. (2019). It takes a village: The role of community size in linguistic regularization. *Cognitive Psychology*, *114*, 101227.

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.

Malouf, R., Ackerman, F., & Semenuks, A. (2020). Lexical databases for computational analyses: A linguistic perspective. *Proceedings of the Society for Computation in Linguistics*, *3*(1), 297–307.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al.. (2020). Unimorph 3.0: Universal morphology.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Sinnemäki, K., & Di Garbo, F. (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.

Spike, M. (2017). Population size, learning, and innovation determine linguistic complexity. In *Cogsci.*

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

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