Investigating relationship between I-complexity and population size

Arturs Semenuks
Department of Cognitive Science, UCSD, San Diego, USA
asemenuk@ucsd.edu

A number of recent theoretical proposals, computational models and cross-linguistic correlational studies suggest that the sociocultural niche a language occupies (exoteric or esoteric) affects the morphosyntactic complexity of that language (Miestamo, 2017). The studies arrive at similar conclusions using approaches differing in methodological details, which could be taken as hinting at the robustness of the hypothesized effect. At the same time, *what* morphological complexity is and what measures best capture it is not yet settled (cf. Berdicevskis et al, 2018). Thus the aforementioned variability in methodological details, e.g. how complexity is operationalized, leaves open some important questions, such as does exotericity cause simplification on *all* dimensions of morphological complexity? Some evidence suggests that the answer might be 'no': 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.

Here I explore the question of whether languages with more exoteric societies, operationalized as having more speakers, tend to have noun paradigms of lower I-complexity, defined by Ackerman and Malouf (2013) as the average conditional entropy between the word forms in paradigms. On the one hand, it should be expected that lower values of I-complexity facilitate learning and would be under stronger selective pressure in languages with more speakers. On the other, Ackerman & Malouf (2013) hypothesize that I-complexity of languages is highly constrained, and thus it might not differ substantially across languages.

I investigate the question using the data for 31 languages from UniMorph 2.0 project containing information about word forms paired with their morphological features and corresponding lemmas (Kirov et al., 2018). 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. No significant correlation between I-complexity and the number of speakers a language is observed in the data (Fig. 1). However, similarly to Ackerman and Malouf (2013), languages are found to have low I-complexities overall, and Monte Carlo simulations show that frequently languages have lower values than expected. The presentation will discuss the possible interpretations of the findings.

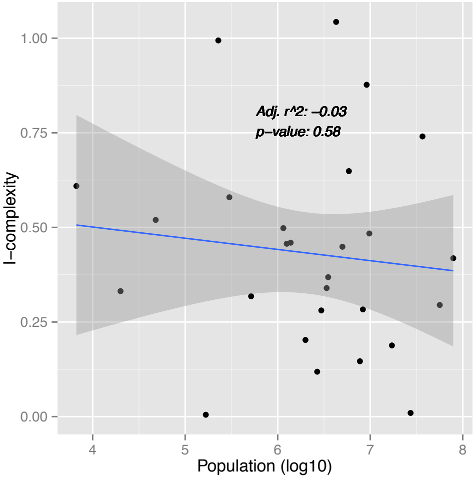

Fig 1*: I-complexity and population size relationship in the data.*

References:

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*, (3), 429-464.

Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., ... & Bentz, C. (2018). Using Universal Dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 8-17.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., ... & Yarowsky, D. (2018). UniMorph 2.0: Universal Morphology. *arXiv preprint arXiv:1810.11101*.

Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue e Linguaggio*, (2), 227–254.

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.