INVESTIGATING CROSS-LINGUISTIC VARIABILITY IN PSYCHOLINGUISTIC DIFFICULTY THROUGH INFORMATION DENSITY VARIABILITY MEASURES

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

Are all languages equally complex? In recent years, this question has attracted attention of researchers from multiple scientific fields (Miestamo, 2017; Ehret, Blumenthal-Dramé, Bentz, & Berdicevskis, 2021). If language is a complex adaptive system (Beckner et al., 2009), we expect languages to differ in their complexity, and a growing number of studies indeed provide support for that position (McWhorter, 2001; Lupyan & Dale, 2010; Trudgill, 2011; Bentz & Winter, 2014; Ehret et al., 2021). This leads to an important question from the perspective of evolutionary linguistics: what leads to the changes in complexity that languages exhibit?

One of the most widely discussed hypotheses attempting to explain (some aspects of) cross-linguistic variability in complexity suggests that languages adapt to the cognitive and sociocultural niches they inhabit, i.e. languages change in complexity to accommodate the cognitive and communicative constraints of their speakers, resulting in, for example, language simplification in communities with more speakers or a higher percentage of non-native learners (Lupyan & Dale, 2010; Bentz & Winter, 2014).

Given this, we can expect a relation between descriptive complexity and psycholinguistic difficulty – less complex languages should be easier to learn, produce and/or comprehend. However, this assumption still requires empirical support, as theory-based complexity and empirically observed cognitive difficulty do not necessarily entail one another (Miestamo, 2017). Surprisingly, the existence of this relation is underresearched. Furthermore, while some studies support this hypothesis (Berdicevskis & Semenuks, 2022), a number of recent studies do not find support for the theorized links between facets of complexity and psycholinguistic difficulty (Atkinson, Smith, & Kirby, 2018; Semenuks & Berdicevskis, 2018; Wagner, Smith, & Culbertson, 2019; Johnson, Gao, Smith, Rabagliati, & Culbertson, 2021). Thus, the question of whether descriptively simpler languages are also easier does not yet have a clear answer, which makes it less clear why

the observed cross-linguistic variability in complexity and the relationships between different complexity dimensions and properties of language speakers and communities exist.

One way to tackle this issue is to use metrics that are more informed by psycholinguistic research and can be reasonably assumed to be transparently related to (or even operationalize) facets of learning, production or comprehension difficulty. A promising candidate for such a metric is the average information density variability of a language at a particular structural level. Information density is the amount of information transmitted per unit at a particular level of organization (e.g. word or syllable), and it can be operationalized as the information content of the unit in its context, i.e. the negative log probability of the unit given its context. For example, in the sentence "I like coffee with milk and sugar", the word "like" has a relatively high information density, as it is relatively unexpected given the preceding context, is more surprising, and has a higher information content, whereas "sugar" has a relatively low information density due to its high predictability (low information content) at the end of the sentence. Theoretical considerations and empirical research suggest that the difficulty of processing a linguistic unit is predicted by its information density. Based on this, Jaeger and Levy (2006) put forward the Uniform Information Density (UID) hypothesis, which argues that speakers aim to minimize the variability in the information density of their utterances, as a UID strategy minimizes the total difficulty of processing an utterance. A variety of studies provide empitical support for the UID hypothesis, e.g. see Genzel and Charniak (2002), Aylett and Turk (2004), Frank and Jaeger (2008).

Thus, we can expect that languages argued to be under higher pressure to be more efficiently structured, such as languages with more speakers, should be more easily processed, and thus have more constrained information density variability of units at different levels of organization. I investigate this hypothesis by calculating the variability in the word information density at the syntactic level using the part-of-speech annotated data for a sample of languages from the Universal Dependencies corpora. Mixed-effects models show that languages vary on this measure and, surprisingly, exhibit higher values of information density variability with more speakers. However, we also find that the average information density level decreases with the number of speakers and is negatively correlated with the information density variability, thus lowering the average unexpectedness of a syntactic unit and creating a tradeoff. Taken together, the results suggest that while individual facets of language complexity correlate with extralinguistic sociocultural properties, they also sometimes trade off with each other, potentially providing a solution to why some previously studied facets of complexity do not correlate with psycholinguistic difficulty.

References

Atkinson, M., Smith, K., & Kirby, S. (2018). Adult learning and language simplification. *Cognitive science*, *42*(8), 2818–2854.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31–56.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al.. (2009). Language is a complex adaptive system: Position paper. *Language learning*, *59*, 1–26.

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

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

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 199–206).

Jaeger, T., & Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*.

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.

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

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

Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue e linguaggio*, *16*(2), 227–254.

Semenuks, A., & Berdicevskis, A. (2018). What makes a grammar difficult? experimental evidence. In *The evolution of language: Proceedings of the 12th international conference (evolangxii).*

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

Wagner, S., Smith, K., & Culbertson, J. (2019). Acquiring agglutinating and fusional languages can be similarly difficult: Evidence from an adaptive tracking study. In *Cogsci* (pp. 3050–3056).