The Equi-complexity vs. Typology: Measurement of Overall Linguistic Complexity and Typological Categories[1]

Takuto NAKAYAMA
Keio University, tnakayama.a5ling@gmail.com

Keywords: linguistic complexity, Shannon entropy, linguistic typology

This research proposes an information theory-based method to measure overall linguistic complexity and demonstrate the similarity of the linguistic complexities of three major typological categories: the agglutinative, fusional, and isolating. Over the last century, most linguists have believed in the equi-complexity of language: "the equi-complexity dogma is that the total complexity of a language is fixed because sub-complexities in linguistic sub-systems trade off. Accordingly, simplicity in some domain A must be compensated by complexity in domain B, and vice versa" (Kortmann and Szmrecsanyi, 2012, p. 7). However, linguists have not yet reached a consensus on how to measure linguistic complexity (Bentz et al., 2022), let alone verified whether the equi-complexity of language is universally applicable. To fully assess the complexity of language, we first need to develop a method for measuring complexity, which is one of the goals of this research.

This research defines language as "a communication between the speaker and hearer" and "a message is 'complex' if it has a large information content" (Joula, 2008, p. 6). This definition is compatible with the concept of Shannon entropy (Shannon, 1948). According to information theory, the complexity of language is high if what unit will appear next in a text is more unpredictable.

The following method is used for the simultaneous measurement of complexity in multiple subdomains of languages (e.g., characters, phonemes, morphemes, words, and constructions). The first step is to calculate the average entropies per unit, which are given by the following formula: $\frac{1}{n}\sum_{i=1,j=n}^{l-n+1,l} p(x_{ij}) log_2 p(x_{ij})$. x_{ij} refers to a string consisting of n units from the ith to the jth one in a text with l units, and $p(x_{ij})$ refers to the probability of how frequently the string appears in the text, from $n = 1$ to the smallest n for which all the strings with n units occur only once in the text. The second step is to obtain the power exponent of the result of the first step regressed on the power law. The first and second steps are done for each subdomain in question, from which we can obtain a value for one subdomain. Then, a vector that consists of these values is given, which simultaneously describes multiple aspects of a text. As Deutscher (2009) stated, a vector form is required to describe overall linguistic complexity. Therefore, one vector is given for one text. The third step is to standardize each value of all the given vectors from the first and second steps into its average of 0 and its variance of 1, as well as to obtain the principle components with Principle Component Analysis (PCA). The advantage is that this method can take into account multiple aspects of language, such as morphological sequence, word sequence, and grammatical sequence.

This pilot study focuses on three languages—English, Japanese, and Chinese—as examples of fusional, agglutinative, and isolating languages, respectively. The data consist of 27 translated excerpts from three versions of the New Testament: the American Standard Version, Kougoyaku (Japanese spoken-style translation), and the Chinese Union Version. The results showed that although the Chinese excerpts reflect some degree of unique behavior, there seems to be no significant difference among those languages, which suggests that the idea of equi-complexity is universally applicable, even among languages with different typological categories.

References

Bentz, Christian., Gutierrez-Vasques, Ximena., Sozinova, Olga., & Samardžić, Tajia. 2022. Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard*.

Deutcher, Guy. 2009. "Overall complexity": A wild goose chase? In Geoffrey Sampson, David Gil, & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 243-251. Oxford: Oxford University Press.

Juola, Patrick. 2008. Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki, & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 89-108. Amsterdam: John Benjamins Publishing Company.

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3). 379-423.

Szmrecsanyi, Benedikt., & Kortmann, Bernd. 2012. Introduction: Linguistic complexity: Second Language Acquisition, indigenization, contact. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic Complexity*, 6-34. Berlin: De Gruyter.

[1] **Acknowledgement**: This work is supported by JST SPRING, Grant Number JPMJSP2123.