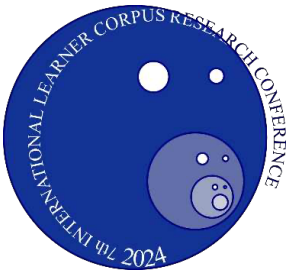
Book of abstracts
LCR2024 Tartu



UNIVERSITY OF TARTU
Institute of Foreign Languages and Cultures

Learner Corpus Association

JOHN BENJAMINS PUBLISHING COMPANY

Complexity or complexities? A simulation study on lexical complexity in expert and learner texts through the lens of information theory

Brasolin, Paolo (independent researcher) and Arianna Bienati (University of Modena and Reggio Emilia / Eurac Research)

The debate about linguistic complexity in general and lexical complexity in particular has been extremely lively, exploring both the validity of indices in capturing the construct (e.g., McCarthy & Jarvis, 2010; Kyle et al., 2021; Zenker and Kyle, 2021) and the theoretical foundations of the construct itself (e.g., Bulté and Housen, 2012; Jarvis, 2013; Pallotti, 2015). Intuitively, complexity transcends "the number and variety of an item's constituent elements" to include "the elaboratedness of their interrelational structure" (Rescher, 2020:1). This echoes the concept of Gell-Mann effective complexity in information theory, which emphasizes the amount of non-random information in a system, which peaks in the intermediate stage between order and disorder. Gell-Mann complexity is often opposed to Kolmogorov complexity, i.e., the total amount of information in a system, which monotonically increases from maximum order to maximum disorder.

This study explores which information-theoretical notion of complexity (Kolmogorov vs. Gell-Mann) is measured by widely used complexity indices, via a simulation study on four Italian corpora, representing the spectrum from expert to learner texts. New texts are synthesized from the originals by altering them in two directions: increased order is obtained as the repetition of increasingly smaller subsections of the original text, whereas increased disorder is obtained as the shuffling of increasingly smaller fragments of it. Additionally, we generate texts with uniform word distribution, simultaneously altering both the structure and the original word distributions. For each corpus, the synthetic data allow us to explore the spectrum from total order to total disorder. All texts are analyzed using type-token-ratio-based and surprisal-based metrics, including fluctuation complexity (Bates and Shepard 1993). Examining the distribution of the computed values shows that TTR-based metrics, except MATTR, are sensitive to increased order but not disorder. Surprisal-based measures, on the other hand, do show interesting Kolmogorov (entropy) or Gell-Mann behavior (normalized entropy and fluctuation complexity), enhancing their mutual interpretability when combined. Our results indicate that fluctuation complexity in particular could complement linguistic complexity tools, since it captures the intuitive notion of complexity in a text.

References

Bates, J.E., & Shepard, H.K. (1993). Measuring complexity using information fluctuation. Physics Letters A, 172(6), 416-425.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. Language Learning, 63(s1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. Language Assessment Quarterly, 18(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Pallotti, G. (2015). A simple view of linguistic complexity. Second Language Research, 31(1), 117-134.

Rescher, N. (2020). Complexity: A Philosophical Overview. Routledge. https://doi.org/10.4324/9780429336591

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. Assessing Writing, 47, 100505. https://doi.org/10.1016/j.asw.2020.100505