Between order and disorder: an information-theoretic approach to linguistic complexity

Abstract

The *surprisal* associated with an event of probability p is defined as $\log(1/p)$. This information-theoretic measure has been proposed and used in both psycholinguistics and corpus linguistics (Gries and Ellis, 2015) to capture the intuitive notion of unexpected events (e.g., words, n-grams, constructions) in language. On the processing side, the tension between the expected and unexpected seems to have important effects on both the processing costs and the activation of mechanisms of explicit learning (Jaeger and Snider, 2008). Regarding production, surprisal-based measures could shed light on the complexity of a user's linguistic system. More proficient language users may not only apply the *normal, conventional usage* of words and constructions, but also exploit them, creating novel, original structures.
In linguistic complexity studies, there have been calls to capture the balance between the expected and the unexpected, conventionality and surprise, order and disorder, structure and lack thereof (Pallotti, 2015). This necessity can be satisfied by the information-theoretical notion of Gell-Mann (effective) complexity, defined as the amount of non-random information in a system: it peaks in the region between order and disorder, where complex structures can manifest. It is often opposed to Kolmogorov (algorithmic) complexity, the total amount of information in a system, which monotonically increases with disorder instead.

In this contribution we focus on lexical complexity to answer the following research question: *which information-theoretical notion of complexity (Kolmogorov vs. Gell-Mann) is measured by the most-used lexical complexity indices?*

To address this question, we conduct a simulation that evaluates lexical complexity on four Italian corpora, which represent the spectrum from expert- to learner-authored texts. We synthesize new texts from the corpora by altering them in two directions: more orderly texts are obtained by repeating increasingly smaller subsections of the text itself until a single word is repeated; whereas a completely disordered configuration is obtained by resampling increasingly smaller n-grams taken from the original text. The synthetic data represents, for each corpus, a way to explore the spectrum from total order to total disorder.

On all texts we compute type-token-ratio-based and surprisal-based metrics, including fluctuation complexity (Bates and Shepard, 1993). A preliminary visual inspection through boxplots shows that TTR-based metrics are sensitive only when there is an increase in order but not in disorder. Surprisal-based measures, on the other hand, do show interesting Kolmogorov (entropy) or Gell-Mann behavior (normalized entropy and fluctuation complexity). Moreover, when used in combination, they mutually enhance their interpretability. Our results indicate that surprisal-based measures could be a useful addition to the toolkit of the linguistic complexity researcher, since they are adherent to both the theoretical and observational underpinnings of usage-based accounts of language learning.

References:

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

Gries, S.T., & Ellis, N.C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning*, *65*(S1), 228-255.

Jaeger, T., & Snider, N. (2008). Implicit Learning and Syntactic Persistence:Surprisal and Cumulativity. CogSci30. Austin:TX.

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

Keywords: complexity theory, lexical complexity, information theory