World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:17, No:07, 2023

Testing the Simplification Hypothesis in Constrained Language Use: An Entropy-Based Approach

Authors : Jiaxin Chen

Abstract : Translations have been labeled as more simplified than non-translations, featuring less diversified and more frequent lexical items and simpler syntactic structures. Such simplified linguistic features have been identified in other bilingualism-influenced language varieties, including non-native and learner language use. Therefore, it has been proposed that translation could be studied within a broader framework of constrained language, and simplification is one of the universal features shared by constrained language varieties due to similar cognitive-physiological and social-interactive constraints. Yet contradicting findings have also been presented. To address this issue, this study intends to adopt Shannon's entropy-based measures to quantify complexity in language use. Entropy measures the level of uncertainty or unpredictability in message content, and it has been adapted in linguistic studies to quantify linguistic variance, including morphological diversity and lexical richness. In this study, the complexity of lexical and syntactic choices will be captured by word-form entropy and pos-form entropy, and a comparison will be made between constrained and non-constrained language use to test the simplification hypothesis. The entropy-based method is employed because it captures both the frequency of linguistic choices and their evenness of distribution, which are unavailable when using traditional indices. Another advantage of the entropy-based measure is that it is reasonably stable across languages and thus allows for a reliable comparison among studies on different language pairs. In terms of the data for the present study, one established (CLOB) and two self-compiled corpora will be used to represent native written English and two constrained varieties (L2 written English and translated English), respectively. Each corpus consists of around 200,000 tokens. Genre (press) and text length (around 2,000 words per text) are comparable across corpora. More specifically, word-form entropy and pos-form entropy will be calculated as indicators of lexical and syntactical complexity, and ANOVA tests will be conducted to explore if there is any corpora effect. It is hypothesized that both L2 written English and translated English have lower entropy compared to non-constrained written English. The similarities and divergences between the two constrained varieties may provide indications of the constraints shared by and peculiar to each variety.

Keywords : constrained language use, entropy-based measures, lexical simplification, syntactical simplification
Conference Title : ICCLTS 2023 : International Conference on Corpus Linguistics and Translation Studies
Conference Location : Rome, Italy
Conference Dates : July 17-18, 2023