keywords: semantic complexity, syntactic complexity, simple English, Wikipedia

The simplification of written English: A study of the Simple English Wikipedia

The problem addressed in this presentation concerns the topic of complexity/simplicity of written English. Specifically, it tackles the issue of simplicity in the Simple English Wikipedia – a version of the most popular online knowledge source designed for people with special educational needs. Issues of formal and/or semantic complexity/simplicity inscribe themselves in research conducted by cognitive linguists.

Two parallel corpora (selected examples of equivalent entries from the "standard" English and the Simple English Wikipedias) have been examined. As a result, seven main simplification mechanisms used by the authors of Simple have been identified: rewording, structure change, deletion, addition, merging, splitting, and content change. It was hypothesized that the mechanisms identified would give an insight into the issue of what makes (written) language complex/simple and, consequently, how to transform what is "complex" into what is "simple."

In my presentation, I will seek an answer to the question whether the mechanisms I distinguished are consistent with traditional views on readability, analyzability, and compositionality as indicators of the level of semantic and structural (morphological and/or syntactic) complexity. Consequently, the presentation will address some problem areas of formal vs. semantic/conceptual complexity/simplicity in language.

Initial findings suggest that there is no universal method of simplifying texts implemented by all the editors of Simple. This may be due to the lack of obligatory Wikipedia guidelines for transforming a standard English text into a "simple" article. However, some tendencies have been observed that can be found in (almost) all entries in Simple. Interestingly, my initial hypothesis that simplifying texts would be based mainly on replacing a "hard" (less frequent, formally and conceptually more complex) word with its more basic equivalent has not been confirmed. As my study shows, word-to-word rewording builds up only a fraction of all the mechanisms observed through corpus analysis.

References:

Coster, W., and Kauchak, D. 2011. "Simple English Wikipedia: A new text simplification task." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*. Portland: Association for Computational Linguistics, 665–669.

Davison, A., and Lutz, D. 1985. "Measuring syntactic complexity relative to discourse context." In: D. Dowty, L. Karttunen, and A. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge: Cambridge University Press.

Goddard, C. 2011. *Semantic Analysis: A Practical Introduction*. Oxford: Oxford University Press.

Siwicki, A. 2017. *Mechanisms involved in written language simplification: A case of Simple English Wikipedia* [unpublished MA thesis]. Warsaw: University of Warsaw.

Word count: 413