On the Correlation of Context-Aware Language Models with the Intelligibility of Polish Target Words to Czech Readers

Klára Jágrová, Marius Mosbach, Michael Hedderich, Tania Avgustinova, Dietrich Klakow
Saarland University, Saarbrücken, Germany
kjagrova@coli.uni-saarland.de

This contribution seeks to provide a rational probabilistic explanation for the intelligibility of words in a genetically related language that is unknown to the reader – a phenomenon referred to as intercomprehension. In this research domain, the intelligibility of stimuli was, among other factors, traditionally explained by linguistic distance and neighbourhood density of the stimulus towards a language in the reader's linguistic repertoire (e.g., Heeringa et al. 2013).

Jágrová & Avgustinova (2019) showed that predictability in context contributes to the intelligibility of the target items. They gathered data from web-based cloze translation experiments for 149 Polish sentences (Block and Baldwin, 2010). These were presented to Czech readers who were asked to translate the highly predictable target words in sentence final position. The majority of the items were more comprehensible within the sentences than if presented without context to another group of Czech respondents. However, for some target words the situation was reversed: the target word intelligibility in context decreased if compared to the condition without context. An error analysis revealed systematic patterns, such as L1/Ln interferences or perceived morphological mismatches. Most of them were in combination with the readers' priming by a dominant concept in the sentence.

Jágrová & Avgustinova (2019) correlated the intelligibility scores of the target words with surprisal values from 3-gram language models (LMs). Since 3-gram surprisal can explain predictability effects only using the two words preceding the target word, the overall correlations with surprisal are low. Interestingly, surprisal correlates stronger with intelligibility of target words that are non-cognates and false friends.

In this contribution we hypothesize that intelligibility of these highly predictable words will correlate better with surprisal values obtained from LMs which incorporate information from the entire sentence. We evaluate two context-aware LM architectures: LSTMs that can take long distance dependencies into account and Transformer based LMs which are able to access the whole input sequence at the same time. We investigate how their use of context affects surprisal and its correlation with intelligibility.

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670. https://doi.org/10.3758/BRM.42.3.665

Heeringa, W., Golubovic, J., Gooskens, C., Schüppert, A., Swarte, F., & Voigt, S. (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In C. Gooskens, & R. van Bezooijen (Eds.), *Phonetics in Europe: Perception and Production* (pp. 99-137). Frankfurt a. M.: P.I.E. - Peter Lang.

Jágrová, K., & Avgustinova, T. (2019). Intelligibility of highly predictable Polish target words in sentences presented to Czech readers. To appear in *Proceedings of CICLing: International Conference on Intelligent Text Processing and Computational Linguistics*. http://www.coli.uni-saarland.de/~tania/ta-pub/CICLing_preprint_Jagrova_Avgustinova_2019.pdf