Tania Avgustinova, Andrea Fischer, Klara Jagrova, Dietrich Klakow, Roland Marti & Irina Stenger
Saarland University
avgustinova@coli.uni-saarland.de, afischer@lsv.uni-saarland.de, kjagrova@coli.uni-saarland.de,
dklakow@lsv.uni-saarland.de, rwmslav@mx.uni-saarland.de, ira.stenger@mx.uni-saarland.de

Orthography in language modeling of mutual intelligibility

While some of the varying aspects and degrees of mutual intelligibility between certain language pairs, such as Czech and Polish, are well-studied in linguistics, systematic large-scale work based on mathematical models and backed by statistical evidence is, overall, still lacking. At the same time, the eye-tracking experiments performed by Smith and Levy (2013) have revealed that surprisal scores assigned by a trigram language model correlate well with reading times and can thus provide reliable estimates for the relative comprehensibility of a text.

In our work we present a statistical language model which can be used to combine linguistic as well as statistical knowledge in one unifying formalism improving upon classic n-gram language models and study its application to modeling cross-lingual intelligibility. This model allows us to study the influence of individual linguistic or statistic features on surprisal and to identify the most important features and combinations thereof.

We focus on the orthographic aspects of Czech-Polish reading intercomprehension. Adaptation between source and target domain is done by deconstructing words into different morphematic representations – derived either automatically from data or designed by trained linguists – and projecting the morphemes of one language into the space of morphemes of the other. To this end, we employ scores derived from variants of Levenshtein distance and similar sequence alignment scores. Varying levels of linguistic inter-language knowledge possessed by the reader are modeled via different morphematic embeddings. The resulting models are tested on a variety of texts, including synthetic texts in which words are replaced with homonyms unknown to the receptive language, in order to show the efficacy of our model.

References

Smith, Nathaniel J., and Roger Levy. 2013. The Effect of Word Predictability on Reading Time Is Logarithmic. In *Cognition* 128.3 (2013). 302–319.