[abstract for poster session]

Can algorithms tell interpreted and non-interpreted speech apart? Looking for grammatical features of Polish interpretese with a Random Forest model

Placed at the crossroads of Translation Studies and Corpus Linguistics, our paper means to contribute to the field of Empirical Translation Studies, particularly within the area of interpreting.

It attempts to do so by interfacing the research paradigm of recurrent features of translation/interpreting and advanced statistical methods that allow to adopt a more data-driven approach, which is still relatively rare in Corpus-based Translation Studies. Following the tradition of investigating interpreted discourse, i.e., interpretese started by Shlesinger (2008), scholars have mostly relied on operationalizations of the phenomena established a priori like the investigations of linguistic patterns based on lexical density, lexical variation, type-token ratio (Sandrelli & Bendazzoli 2005). Studies exploring cohesiveness or explicitness in interpreting looked at connectives and cohesive devices (Defrancq, Plevoets & Magnifico 2015).

In this paper, we apply a data-driven approach to investigate the grammatical features of Polish interpretese in the Polish Interpreting Corpus (PINC) comprising transcriptions of English-Polish and Polish-English interpretations of the speeches delivered at the plenaries of the European Parliament and their source texts. Following, in part, the methodology adopted in the study of constrained written varieties by Ivaska and Bernardini (2020) we first apply the Boruta feature selection method (Kursa & Rudnicki 2010) to retrieve a set of POS bigrams that allow to differentiate between interpreted and non-interpreted language best. Then we train a Random Forest model to establish the accuracy with which the selected features allow the model to tell interpreted and non-interpreted Polish discourse apart. Our preliminary results show that the POS bigrams help distinguish the two datasets with 0.7368 accuracy. Features that contribute to the outcomes most include, among other, the following POS bigrams sconj_verb, noun_sconj, noun_noun, noun_adj, intj_verb, verb_part, part_verb, adp_verb. These sequences are further analysed to narrow down the list only to the features characteristic of the Polish interpretese.

Corpus-based Interpreting Studies have so far lacked contributions on Polish interpretese. This paper aims to fill this gap by revealing differences in grammatical structures in Polish used by interpreters and non-interpreters. We will look into overused and underused POS bigrams hoping to discover how formulating and linking ideas in interpreting differs from unconstrained communicative situation due to bilingual processing and text dependent production constraints.

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Yearbook of Corpus Linguistics and Pragmatics 2015, 195–222. Switzerland: Springer.

Ivaska, Ilmari & Silvia Bernardini. 2020. Constrained language use in Finnish: A corpus-driven approach. Nordic Journal of Linguistics 43(1). 33–57.

Kursa, Miron B. & Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. Journal of Statistical Software 36. 1–13. https://doi.org/10.18637/jss.v036.i11.

Sandrelli, Annalisa & Claudio Bendazzoli. 2005. Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus). In Proceedings from the Corpus Linguistics Conference Series, vol. 1.

Shlesinger, Miriam. 2008. Towards a definition of Interpretese An intermodal, corpus-based study. In Gyde Hansen, Andrew Chesterman & Heidrun Gerzymisch-Arbogast (eds.), Efforts and models in interpreting and translation research: A tribute to Daniel Gile, vol. 80, 237.

[364 words]