Contrastive Analysis of Articles in Romance Languages and Croatian on a Parallel Corpus

In the first part of our talk, we will present the construction of a singular new resource. *RomCro* is a parallel multilingual and multidirectional corpus of five Romance languages (Spanish, French, Italian, Portuguese, Romanian) and Croatian. The corpus, counting 15.9 million words, contains original sentences from literary texts from the 20th and 21st centuries, aligned with their translational equivalents in the remaining languages. Since the original sentence order has been scrambled, the corpus is available for non-commercial use under the license CC-BY-NC-4.0 via platforms *Sketch Engine* and ELRC. This type of corpora has a wide use and is known for its application in different kinds of linguistic research (contrastive linguistics, translation studies, phraseology, lexicography, etc.) (e. g. Granger, Lerot & Petch-Tyson, 2003; Teubert, 2007), translation training (López Rodríguez, 2016) and training of machine translation systems (Koehn et al., 2007), as well as terminology extraction (Lefever, Macken & Hoste, 2009).

In the second part, we would like to point out some of our research based on the data extracted from *RomCro*. We investigated similarities and differences between five Romance languages in the use of definite and indefinite articles (including the absence of article, i. e. zero article). Although most of the grammatical rules are similar in all Romance languages (especially when it comes to the definite article) (Academia Română, 2008; Buzaglo Paiva Raposo et al., 2013; Grevisse & Goosse, 2008; Real Academia Española, 2009; Enciclopedia dell'Italiano), we observed some interesting differences (e. g. more common use of possessive in French instead of definite article or higher frequency of zero article in the so called peripheric Romance languages). However, what especially attracted our attention is the possibility of switching from definite to indefinite article and vice versa regardless of the language (we call it "change of perspective") (Autor, 2020). We explain that feature by the fact (in our opinion, usually overlooked) that a noun in discourse can carry various characteristics regarding its determination and it is up to the author (or, in this case, translator) to choose which one they will point out. It is interesting to notice that sometimes the translator does not follow the author's choice (regardless of the fact that such a possibility exists in the language).

We think that the use of *RomCro* can prove itself to be very useful in contrastive Romance linguistics, but also in the comparison of Romance languages and Croatian, a Slavic language that does not have articles as a morphological category. Our next aim is to find out the possible differences in the translation of a Croatian original to the Romance languages regarding the expression of the noun determination.

Keywords: multilingual corpus, parallel corpus, Romance languages, Croatian, article.

References

Academia Română. 2008. *Gramatica limbii române*. Editura Academiei Române, Bucureşti.

Buzaglo Paiva Raposo, E. et al. 2013. *Gramática do português*. Fundação Calouste Gulbenkian, Lisboa.

Enciclopedia dell'Italiano. https://www.treccani.it/enciclopedia.

Granger, S., J. Lerot, & S. Petch-Tyson (eds.). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Rodopi, New York.

Grevisse, M. & A. Goosse. 2008. *Le bon usage*. Éditions De Boeck Université, Bruxelles.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177-180, Association for Computational Linguistics.

Lefever, E., L. Macken, & V. Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 496-504, Athens.

López Rodríguez, C. I. 2016. Using Corpora in Scientific and Technical Translation Training: Resources to Identify Conventionality and Promote Creativity. *Cadernos de traduçao*, 1:88-120.

Real Academia Española & Asociación de Academias de la Lengua Española. 2009. *Nueva gramática de la lengua española*. Espasa Libros, Madrid.

Teubert, W. (ed.). 2007. *Text Corpora and Multilingual Lexicography.* John Benjamins Publishing Company, Amsterdam / Philadelphia.

Autor. 2020.