Simon Krek/Polona Gantar/Iztok Kosem

EXTRACTION OF COLLOCATIONS FROM THE GIGAFIDA 2.1 CORPUS OF SLOVENE

Abstract This paper describes a method for extracting collocation data from text corpora based on a formal definition of syntactic structures, which takes into account not only the POS-tagging level of annotation but also syntactic parsing (syntactic treebank model) and introduces the possibility of controlling the canonical form of extracted collocations based on statistical data on forms with different properties in the corpus. Specifically, we describe the results of extraction from the syntactically tagged Gigafida 2.1 corpus. Using the new method, 4,002,918 collocation candidates in 81 syntactic structures were extracted. We evaluate the extracted data sample in more detail, mainly in relation to properties that affect the extraction of canonical forms: definiteness in adjectival collocations, grammatical number in noun collocations, comparison in adjectival and adverbial collocations, and letter case (uppercase and lowercase) in canonical forms. The conclusion highlights the potential of the methodology used for the grammatical description of collocation and phrasal syntax and the possibilities for improving the model in the process of compilation of a digital dictionary database for Slovene.

Keywords Collocations; discovering collocations in corpora; digital collocation database

Contact Information

Polona Gantar
University of Ljubljana
apolonija.gantar@guest.arnes.si

Iztok Kosem
Jožef Stefan Institute & Faculty of Arts, University of Ljubljana
iztok.kosem@ijs.si

Simon Krek
Jožef Stefan Institute
simon.krek@guest.arnes.si

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.