Polona Gantar/Simon Krek

CREATING THE LEXICON OF MULTI-WORD EXPRESSIONS FOR SLOVENE

Methodology and structure

Abstract This paper describes a method for automatic identification of sentences in the Gigafida corpus containing multi-word expressions (MWEs) from the list of 5,242 phraseological units, which was developed on the basis of several existing open-access lexical resources for Slovene. The method is based on a definition of MWEs which includes information on two levels of corpus annotation: syntax (dependency parsing) and morphology (POS tagging), together with some additional statistical parameters. The resulting lexicon contains 12,358 sentences containing MWEs extracted from the corpus. The extracted sentences were analysed from the lexicographic point of view with the aim of establishing canonical forms of MWEs and semantic relations between them in terms of variation, synonymy, and antonymy.

Keywords Identification of multi-word expressions; multi-word expressions lexicon; canonical form of multi-word expressions; Slovene; digital dictionary database

Contact Information

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