EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Book of Abstracts

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

The XIX EURALEX International Congress: Lexicography for inclusion

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian

Iztok Kosem[1,2], Simon Krek[2], Polona Gantar[1]

[1] *Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*
[2] *Jožef Stefan Institute, Ljubljana, Slovenia*

Abstract

Finding a way to manage large interconnected datasets, and not viewing different lexicographic resources as isolated units, has become a very important challenge in modern lexicography. Namely, one quickly faces considerable duplication of work if projects with different lexicographic focus, such as general dictionaries, collocations dictionaries, thesauri and so forth, are compiled completely separately, by different teams, even if several parts of microstructure are shared. Of course, target users can be different, requiring different definitions, examples etc.; however, this should not be used as an excuse why a common database of integrated resources has not been considered.

There is already evidence about the topicality of such approach to data modelling across Europe, as several institutions have presented plans of data consolidation, for example for Estonian (Tavast et al. 2018), German (Geyken 2019), Polish (Żmigrodzki 2018), and Dutch (Colman 2016). Large databases of a variety of data on a language have also been a topic at a workshop on the Future of Academic Lexicography, held in Leiden in November 2019. A further boost to this trend will be provided by tools produced as part of the ELEXIS infrastructure, which will enable linking of lexicographic resources intralingually and cross-lingually.

In this paper, we present the case of the Digital Dictionary Database for Slovenian, which aims to become a one-for-all database for the Slovenian language, to be used for both in the compilation of language resources and natural language processing tasks. The plans for the database have been described in detail in Klemenc et al. (2017). At the time of writing, the database contained data from two resources: Slovene morphological lexicon and the Collocations Dictionary of Modern Slovene. But with several other projects ongoing, the database modelling is continuous and the model is therefore regularly being updated to cater for different types of data, such as synonyms (from the Thesaurus of Modern Slovene), translations (from the bilingual dictionaries, currently Slovenian-Hungarian dictionary), and corpus examples (authentic or modified ones). Consequently, we can present only current status of the model, knowing that it will soon undergo more changes. First, we will look at the database model, focussing on the most challenging concepts, for example lexeme and lexical unit, and connections between different constituent parts of the model. Relatedly, the schema used for exporting data from the database into external tools such as dictionary writing systems is examined; several decisions that ensure compatibility between data models of different lexicographic resources had to be made, and we will discuss both advantages and shortcomings.

The second part of the paper looks in more detail at the ontology of semantic types that we devised for our lemmas, and also collocations; we believe that we will later be able to apply these semantic types to multiword units and patterns (or different arguments in patterns). Semantic types are closely related to our database model, as they represent a type of information we will record to enable linking our data with other languages. We have tested different ontologies, for example Wordnet, Corpus Pattern Analysis (CPA), Lexiconet, Framenet, Estonian ontology, and Simple-Clips ontology, and established that none of them completely met our purpose. For example, Wordnet lexicographer files were too broad, while CPA categories and especially Lexiconet and Framenet categories were in places too fine-grained and/or too overlapping. Still, it did not make sense to start completely from scratch, so we took Wordnet categories and tried to divide them to the extent where each subcategory had enough

representatives to legitimize its inclusion. At the time of writing, the ontology is still being finalised, but it is known that there are 19 top-level categories which are almost overlapping with Wordnet categories – this is not a coincidence as it was our aim to make linking our database with databases of other languages as straightforward as possible. More significant changes can be observed at subcategory levels, where in many categories one can find many similarities with Lexiconet, but without the repetitiveness of possible categories (e.g. Lexiconet groups substances by source, by function and by natural state, whereas our categorization follows a more source-function division).
Further plans include analysing the relationship between our semantic types and semantic concepts of words, in order to determine whether direct mapping is possible, or to what extent. We also plan to examine to what degree do methods such as word sense induction succeed in grouping collocations that belong to the same semantic type. We will also speak briefly about the potential of semantic types for dictionary users, as such information, which might not even be made visible, can facilitate many useful searches in the dictionary, without necessitating any changes to the key parts of the dictionary microstructure.
Bringing it all together in the end, we aim to demonstrate how all this, i.e. semantic types and interconnected databases, can be efficiently used to improve everyday lexicographic workflow, and can facilitate the compilation of new resources, and linking with other existing foreign resources.

Keywords: digital dictionary database, Slovenian, collocations, semantic types, semantic concepts

References

Arhar Holdt, Špela; Čibej, Jaka; Dobrovoljc, Kaja; Gantar, Apolonija; Gorjanc, Vojko; Klemenc, Bojan; Kosem, Iztok; Krek, Simon; Laskowski, Cyprian; Robnik Šikonja, Marko. (2018). Thesaurus of Modern Slovene: By the Community for the Community. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Colman, Lut. (2016). Sustainable lexicography: where to go from here with the ANW (Algemeen Nederlands Woordenboek, an online general language dictionary of contemporary Dutch)? *International Journal of Lexicography*, 29/2, pp. 139-155.

Geyken, Alexander. (2019). The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. Iztok Kosem & Tanara Zingano Kuhn (eds.) *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography. Book of abstracts*. Lexical Computing CZ s.r.o., Brno, Czech Republic.

Klemenc, Bojan; Robnik-Šikonja, Marko; Fürst, Luka; Bohak, Ciril; Krek, Simon. (2017). Technological Design of a State-of-the-art Digital Dictionary. Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok, Krek, Simon (eds). *Dictionary of modern Slovene: problems and solutions.* 1st ed. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.

Kosem, Iztok; Krek, Simon; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka; Laskowski, Cyprian Adam. (2018). Collocations dictionary of modern Slovene. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 989-997. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 749-761.

Tiberius, Carole; Niestadt, Jan. (2010). The ANW: an online Dutch Dictionary. Anne Dykstra & Tanneke Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Ljouwert: Fryske Akademy/Afûk, 181 (abstract).

Żmigrodzki, Piotr. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 209-219. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/118/211/2973-1?inline=1