Vasyl Starko

USL: A COGNITIVELY INSPIRED LEXICON FOR SEMANTIC TAGGING

Keywords Semantic tagging; semantic lexicon; Ukrainian; corpus annotation; cognitive semantics

Semantics is a language level that is extremely important and yet notoriously difficult for natural language processing. Approaches to semantic annotation have been dominated by lexicon-based solutions such as FrameNet and WordNet (Piao et al. 2015). A significant number of corpora are semantically annotated using a version of the WordNet for the language in question. A potential weakness of the WordNet is the low level of granularity, which may complicate downstream tasks. At least for some tasks, a more coarse-grained classification scheme based on the so-called supersenses has been utilized to reduce the average number of senses per word (Ciaramita/Altun 2006).

A different type of a semantic lexical resource for semantic tagging is exemplified by the USAS semantic lexicon (Rayson et al. 2004), which assigns semantic tags based on a universal annotation scheme. However, its dichotomous classification scheme can at times be too rigid, while the top layers of its hierarchy may be excessively abstract for a non-specialist user.

Keeping this in mind, a relatively coarse-grained semantic lexicon has been created based on a classification scheme that reflects the types of categories used by ordinary speakers in speech acts. This lexicographic resource, titled the Ukrainian Semantic Lexicon, has been developed for the Ukrainian language with a possibility of extension to other languages. The USL is a machine-readable dictionary in which each lemma (more precisely, each sense of a given lemma) is supplied with a string of semantic tags. For example,

naukovets 'scholar' 1:conc:hum:prof,

where **1** is the number of the sense, the **conc** tag means 'concrete noun', **hum** 'human being', and **prof** 'profession'. Another example illustrates semantic tags for an adjective:

velykyi 'large, great' 1:size:2:degree:3:age,

where the second size refers to such contexts as *velyka radist* 'great joy'.

Three more classes of words—verbs, adverbs, and numerals—are represented in the USL:

stoiaty 'to stand' 1:loc:body:noncaus:2:loc:noncaus,

where **loc:body** refers to a body position, **loc** expressed the idea of location in general, and **noncaus** points to the noncausative nature of the verb's senses.

povnistiu 'completely' 1:physqual:2:degree:max,

where the Ukrainian adverbs mirrors two senses of its English equivalent.

Numerals in the USL include both quantifiers and absolute quantities:

bahato 'many, a lot' 1:quantif

simdesiat 'seventy' 1:abst:quantity:absol

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.

A cognitive linguistic approach has been adopted to develop a system of semantic classification for the USL. In the center of attention are the so-called basic-level categories, which enjoy a privileged status in natural language categorization and are characterized by a convergence of perceptual, behavioral, and abstract features (Taylor 1995). From this level, relations may extend up and down but only to a limited degree, allowing for shallow hierarchies. In general, the semantic classification in USL is based on a faceted, rather than hierarchical, approach and multiclass membership is possible.

The Ukrainian Semantic Lexicon is now in its second iteration (USL 2.0) and contains 80,000 entries. It is suitable for NLP applications and, indeed, has been used in conjunction with the TagText tagger that performs both morphological and semantic annotation of Ukrainian texts. Both resources are available to lexicographers, computational linguists, and NLP researchers from their respective github repositories (Rysin 2022), (Ukrainian Semantic Lexicon). Both tools have been utilized to semantically tag a large reference corpus of Ukrainian – the General Regionally Annotated Corpus of Ukrainian (GRAC). The full classification scheme is presented on GRAC's website (Shvedova et al. 2017–2022), while a discussion of the theoretical foundations is provided in Starko (2020) and (2021).

In the future, the USL will be further expanded, fine-tuned, and applied in projects involving semantic tagging.

References

Ciaramita, M./Altun, Y. (2006): Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Jurafsky, D./Gaussiere, E. (eds.): Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP). Sydney, pp. 594–602.

Piao, S./Bianchi, F./Dayrell, C./D'Egidio, A./Rayson, P. (2015): Development of the multilingual semantic annotation system. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2015). Denver, pp. 1268–1274.

Rayson, P./Archer, D./Piao, S./McEnery, T. (2004): The UCREL semantic analysis system. In: Proceedings of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks. Lisbon, pp. 7–12.

Rysin, A. (2022): LanguageTool API NLP UK Project. https://github.com/brown-uk/nlp_uk (last access: 25-03-2022).

Shvedova, M./von Waldenfels, R./Yarygin, S./Rysin, A./Starko, V./Nikolajenko, T. (2017–2022): GRAC: General Regionally Annotated Corpus of Ukrainian. Kyiv/Lviv/Jena. http://uacorpus.org/ (last access: 25-03-2022).

Starko, V. (2021): Implementing semantic annotation in a Ukrainian corpus. In: Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), April 22–23, Kharkiv, Ukraine. Volume I: Main conference. Kharkiv, pp. 435–447.

Starko, V. (2020): Semantic annotation for Ukrainian: categorization scheme, principles, and tools. In: Procedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), April 23–24, Lviv, Ukraine. Volume I: Main conference. Lviv, pp. 239–248.

Taylor, J. R. (1995): Linguistic categorizaton. 2nd edition. Oxford.

Ukrainian Semantic Lexicon. https://github.com/brown-uk/dict_uk/tree/master/data/sem (last access: 25-03-2022).

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.

Contact information

Vasyl Starko
Ukrainian Catholic University
v.starko@ucu.edu.ua

Acknowledgements

This project has been supported by a grant from the Believe in Yourself Foundation at the Ukrainian Catholic University in Lviv.

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.