The ELEXIS parallel sense-annotated corpus

Simon Krek[1], Carole Tiberius[2], Kaja Dobrovoljc[1], Jaka Čibej[1], Polona Gantar[3], Jelena Kallas[4], Kristina Koppel[4], Svetla Koeva[5], Veronika Lipp[6], László Simon[6]

[1]Jožef Stefan Institute, Slovenia, [2]Instituut voor de Nederlandse Taal, The Netherlands, [3]Faculty of Arts, University of Ljubljana, Slovenia, [4]Institute of the Estonian Language, Estonia, [5]Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria, [6]Hungarian Research Centre for Linguistics, Hungary

What is it?

An entirely manually-curated lexical-semantic resource available in ten European languages combining corpora and sense inventories

Corpora

- Origin: WikiMatrix
- Number of sentences per language: 2,024
- Identical sentences in 10 languages
- Manually checked translations
- Manually checked annotations

Annotation layers

- tokenisation
- sub-tokenisation
- lemmatisation
- part-of-speech tagging (UD)
- word sense disambiguation

Languages, tokens, lemmas, content words

Language	Tokens	Unique Lemmas	Nouns	Verbs	Adjs	Advs
Bulgarian	33,994	6,683	7,892	3,970	3,313	1,157
Danish	32,524	6,832	7,322	3,099	2,626	1,677
Dutch	34,923	6,488	7,142	3,004	2,833	1,020
English	34,228	6,297	6,716	2,946	2,818	1,079
Estonian	37,693	6,112	8,189	3,327	2,310	1,487
Hungarian	29,657	7,457	6,930	2,485	3,561	1,173
Italian	39,067	6,371	7,864	3,022	2,961	1,368
Portuguese	38,723	6,260	7,372	3,181	2,757	1,302
Slovene	31,237	6,688	7,550	2,579	3,820	1,077
Spanish	37,693	6,112	8,189	2,806	3,141	1,140

Sense inventories (dictionaries, WordNets)

- The Dictionary of Modern Bulgarian
- DanNet (Danish WordNet)
- Open Dutch WordNet
- English WordNet
- EKI combined Dictionary (Estonian)
- The Explanatory Dictionary of the Hungarian Language
- PAROLE-SIMPLE-CLIPS + ItalWordNet (Italian)
- Dictionary of the Lisbon Academy of Sciences (Portuguese)
- Digital Dictionary Database for Slovene
- Spanish Wiktionary

Availability (repository & online concordancer)

Martelli, Federico, et al. 2022. **Parallel sense-annotated corpus ELEXIS-WSD 1.0**, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042.

ELEXIS WSD Dataset and UniDive

Relevance: **WG1** and **WG2**

Possible extension of the current dataset with additional languages and additional annotation layers

- annotation of multiword expressions following the PARSEME annotation guidelines
- annotation of named entities
- syntactic parse structure following Universal Dependencies

Design of the dataset

Martelli, Federico, et al. 2021. **Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages**. In Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 377-395.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.