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

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

Manually-curated lexical-semantic resource combining corpora and sense inventories

Corpora and sense inventories

- Corpora
  - Origin: WikiMatrix
  - Sentences per language: 2,024
  - Identical sentences in 10 languages
  - Manually checked translations
  - Manually checked annotations
- Sense inventories
  - Dictionaries: Bulgarian, Estonian, Hungarian, Portuguese, Slovenian, Spanish
  - WordNets: Danish, Dutch, English, Italian

ELEXIS WSD dataset & 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

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