The ELEXIS parallel sense-annotated corpus

Simon Krek[1], Carole Tiberius[2], Kaja Dobrovoljc[1], Jaka Čibej, 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

{simon.krek,kaja.dobrovoljc,jaka.cibej}@ijs.si,
carole.tiberius@ivdnt.org, apolonija.gantar@guest.arnes.si,
{jelena.kallas,kristina.koppel}@eki.ee,
{lipp.veronika,simon.laszlo}@nytud.hu

Relevant UniDive working groups: WG1, WG2

1 Introduction

A major limitation affecting many research fields including lexicography and NLP is the lack of high-quality manually-curated data which is labour-intensive and costly to produce. Fortunately, recent advances in NLP have shown their effectiveness for the creation and analysis of lexical-semantic resources both within and across languages. However, we believe that such approaches should be a starting point and that robust lexical-semantic studies should rely on manually-curated data whenever possible, which would also encourage deeper connections between lexicography and NLP, for example.

In this context, a parallel sense-annotated corpus has been produced within the H2020 ELEXIS[1] project, an entirely manually-curated lexical-semantic resource available in 10 European languages - Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene and Spanish - featuring 5 annotation layers, i.e. tokenisation, sub-tokenisation, lemmatisation, part-of-speech tagging and word sense disambiguation (see Martelli et al. 2021, 2022).

2 Corpus compilation

The corpus was compiled by automatically extracting a set of sentences from WikiMatrix (Schwenk et al., 2019), a large open-access collection of parallel sentences derived from Wikipedia, using an automatic approach based on multilingual sentence embeddings. The sentences were manually validated according to specific formal, lexical and semantic criteria (e.g. by removing incorrect punctuation, morphological errors, notes in square brackets and etymological information typically provided in Wikipedia pages). To obtain a satisfying level of semantic coverage, sentences with less than 5 words and fewer than 2 polysemous words were filtered out. Subsequently, in order to obtain datasets in the other nine target languages, for each selected sentence in English, the corresponding WikiMatrix translation into each of the other languages was retrieved. Next, these translations were manually checked and missing translations were added, resulting in the same 2,024 sentences for each of the 10 languages.

Once the corpus data was ready, the corpus was automatically tokenised, tagged and lemmatised in accordance with the cross-linguistically harmonised Universal Dependencies (UD) annotation scheme. By default the UDPipe v2.6[2] pipeline was used, but some languages used their own pipeline, e.g. Bulgarian (see Martelli et al. 2021). The results from the automatic processing were again manually checked and adjusted, after which sense annotations were added using a special annotation platform LexTag.

3 Semantic Annotation

In the semantic annotation phase, the tokens in the datasets were manually annotated as either content-words (and assigned a sense) or non-content words (with no assigned sense). The

[1] https://elex.is/

[2] https://lindat.mff.cuni.cz/services/udpipe/

dataset contains between 14,000 and 18,000 content words for each language that have been assigned a sense from the selected sense inventory (i.e. a dictionary) for that language or from BabelNet. Only sense inventories with a Creative Commons licence could be selected. The senses were mostly assigned from the original sense inventories selected for each language (between 11,000 and 15,000 words), with a smaller percentage annotated with senses from BabelNet (up to 2,600 for Italian and as little as 2 for Dutch). Due to the specific nature of the data set (including sentences from specialised domains, not all lemmas were present in the selected sense inventories. Therefore new senses could be added manually during annotation: up to approx. 5,900 for Danish and as little as 119 for English). The extent of adding new senses was left to the discretion of individual language teams depending on the structure of their selected sense inventories and their goals. The number of unannotated tokens also varies between languages. The unannotated tokens are mostly named entities; while some language teams added named entities to their sense inventories as separate senses (e.g. Bulgarian and Hungarian), others decided to process named entities at a later stage.

4 Conclusion and further work

The corpus has been made available through the CLARIN.SI repository (Martelli et al. 2022) in a CONLL-like tab-separated format. In order, the columns contain the token ID, its form, its lemma, its UPOS-tag, its whitespace information (whether the token is followed by a whitespace or not), the ID of the sense assigned to the token, and the index of the multiword expression (if the token is part of an annotated multiword expression). The corpus has also been uploaded to the noSKE[3] online concordancer for easy exploration by linguists with less technical skills.

In the context of UniDive, we would like to extend the current dataset with additional languages and additional annotation layers, e.g.

- Annotation of MWEs including VMWE following the PARSEME annotation guidelines[4]. Although it was possible to annotate MWEs, MWE annotation has not been performed systematically across all 10 languages.
- Annotation of named entities. During the project, named entities were annotated in the English data set which could be used as a reference for annotation in the other languages. However, the subsequent annotation of named entities was left to the discretion of the individual language teams for all languages.
- Syntactic parse structure following UD.

Annotated corpora constitute the Action's major operational tools for NLP-applied universality *(WG1)*. The resulting ELEXIS parallel sense-annotated corpus can be used to perform lexical-semantic analysis across languages, but it can also be used to carry out performance evaluation in both supervised (cf. Barba et al. 2021) and knowledge-based WSD approaches (cf. Maru et al. 2019). The envisaged additional annotation layers can provide a starting point for categorising non-verbal MWEs and for recognising (multi-word) named entities to the (multi-word) concepts *(WG2)*.

References

Barba, Edoardo, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole

[3] Available at CLARIN.SI NoSketchEngine: https://www.clarin.si/noske/run.cgi/corp_info?corpname=elexiswsd&struct_attr_stats=1

[4] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej and Tina Munda 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.

Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej and Tina Munda. 2022. Parallel sense-annotated corpus ELEXIS-WSD 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1674

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3534–3540, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.