A Proposal for a Lexicon-Corpus Interface: The Interplay Between a Concordancer and a Collaborative Knowledge Base Management Tool

Jaka Čibej
University of Ljubljana
Slovenia
jaka.cibej@ff.uni-lj.si

Simon Krek
University of Ljubljana
Slovenia
simon.krek@ff.uni-lj.si

Carole Tiberius
Dutch Language Institute
The Netherlands
carole.tiberius@ivdnt.org

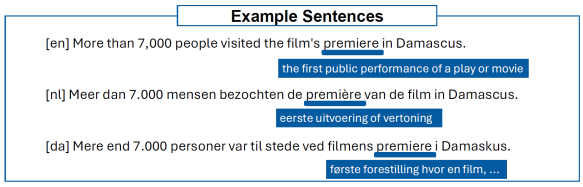

Figure 1: Examples from the *ELEXIS-WSD* Parallel Sense-Annotated Corpus.

Relevant UniDive working groups: WG2

1 Introduction

Within the UniDive COST Action (Savary et al., 2024), one of the tasks within WG2 is the design of a lexicon-corpus interface, i.e. a solution that enables interlinking lexicon entries with their occurrences in corpora. Digital lexicons are complementary to corpora because they aim at holistic language modeling and potentially describe a very wide range of linguistic objects, whereas in corpora many phenomena occur rarely or never.

In this paper, we present one of the outcomes of T2.2 within WG2: a proposal for a possible infrastructure that links a corpus and a lexicon, but also promotes a paradigm shift from lexeme-focused to sense-focused links to corpora. The infrastructure consists of a sense-annotated corpus, a knowledge base management tool, and a concordancer, as presented in the following sections.

2 Sense-Annotated Corpus

The *ELEXIS-WSD* Parallel Sense-Annotated Corpus (Čibej et al., 2025; described in more detail by Martelli et al., 2021) consists of (1) subcorpora for different languages, and (2) sense inventories for each language (i.e. a list of lexemes and their possible senses). The tokens in the corpus are annotated with senses from the sense inventory as shown in Figure 1.

Figure 2: Basic properties of the lexeme 'war' on Wikibase.

3 Knowledge Base Management Platform

The sense inventory of the sense-annotated corpus was uploaded to a knowledge base management platform. In our case, we used *Wikibase Cloud*[1], which enables the storage of structured data (like a databases or knowledge graphs). The data is organized in the form of combinations of *items* and *properties* (e.g., *Paris* (item) → *capital of* (property) → *France* (item)). Each lexeme from the uploaded sense inventory (e.g., *war*, noun) is treated as a separate item with basic features (such as ID, language, and lexical category), as shown in Figure 2. Individual senses, their IDs and definitions are listed below the lexeme as shown in Figure 3 (in the *Senses* section). Each sense can be assigned additional *statements*. In our case, the sense contains two statements: the source ID (the original sense ID from the sense inventory) and the link to *ELEXIS-WSD*, which leads directly to the corpus in a concordancer (see Section 4).

[1]Wikibase Cloud: https://www.wikibase.cloud/

Figure 3: Properties and statements for one of the senses of the lexeme ‘war’ on Wikibase.

4 Concordancer

We used the *NoSketch Engine*[2] concordancer in our proposal. The *ELEXIS-WSD* corpus was converted from CoNLL-U[3] to VERT.[4] Each assigned definition and sense ID from the sense inventory (see Figure 1) were added as attributes to each annotated token in the VERT file, which enables users to query the corpus for tokens annotated with specific senses using Corpus Query Language (CQL).[5]

5 Linking the Corpus and the Lexicon

CQL links to the corpus were generated for each sense in the sense inventory and added as *statements* underneath each definition within *Wikibase Cloud*. For instance, Figure 3) shows sense *L4232-S1* for the English lexeme ‘war’ (noun), with the definition ‘the waging of armed conflict against an enemy’. The same lexeme contains three additional senses with different definitions – each was assigned a different link to lead only to occurrences of the lexeme with the relevant sense. An example of concordances from the English subcorpus of *ELEXIS-WSD* in *NoSketch Engine* containing occurrences of ‘war’ annotated as ‘the waging of armed conflict against an enemy’ is shown in Figure 4.

6 Conclusion and Future Work

We have presented a proposal on how to organize corpus and lexicon data in a linked way, moving from lexeme-based to more sense-focused links to corpora on the one hand, and encoding lexicon data (particularly semantic data) in a structured machine-readable format that can be useful for the compilation of training or evaluation datasets. Although our proposal is just a prototype using existing available resources and platforms, it nevertheless provides a good starting point for future developments.

First, the method relies on a sense-annotated corpus. *ELEXIS-WSD* was annotated manually, but modern methods using large language models have been shown to be useful for word sense disambiguation and annotation of larger amounts of data (Stanković et al., 2026).

Second, a local dedicated installation of *Wikibase* is required for large-scale projects, as the publicly available installation is characterized by size restrictions.

Third, links between sense inventories on the one hand and subcorpora on the other can be extended by also linking identical or similar senses between sense inventories of different languages, further increasing the interconnectedness of the dataset. This is particularly useful for advanced lexicographic resources and databases that can be used to form knowledge graphs for fine-tuning and training LLMs. This would allow for further cross-linguistic comparisons and cross-lingual sense-based queries in corpora. For instance, additional statements could be added to Wikibase underneath each sense, leading to occurrences in corpora of a different language annotated with the same sense.

The framework designed within WG2 of UniDive will be expanded upon within the upcoming *ELEXAI* project (*European Lexicographic Infrastructure for Artificial Intelligence*; 2026-2029), the successor to the *ELEXIS* project (*European Lexicographic Infrastructure, Horizon 2020, No 731015.*),[6] which will further explore the potential of parallel semantic data for the advancement of lexicographic resources and infrastructure with the help of artificial intelligence. In addition, the question of links to existing similar resources such as *Wikidata*[7] or *BabelNet*[8] will be elaborated in order to avoid redundancy and foster the integration of *ELEXIS-WSD* into existing workflows.

[2]About *NoSketch Engine*: https://www.sketchengine.eu/nosketch-engine/

[3]CoNLL-U: https://universaldependencies.org/format.html

[4]VERT (vertical file): https://www.sketchengine.eu/glossary/vertical-file/

[5]Corpus Query Language: https://www.sketchengine.eu/documentation/corpus-querying/

[6]About *ELEXIS*: https://project.elex.is/

[7]Wikihttps://www.wikidata.org/

[8]BabelNet: https://babelnet.org/

Figure 4: Examples containing the lexeme 'war' defined as 'the waging of armed conflict against an enemy' as shown in NoSketchEngine.

Acknowledgements

The presented work was supported by the COST Action CA21167 – *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive) and the research programme *Language Resources and Technologies for Slovene* (P6-0411) funded by the Slovenian Research and Innovation Agency.

References

Jaka Čibej, Simon Krek, Carole Tiberius, Federico Martelli, Roberto Navigli, 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, Rob Tempelaars, Rute Costa, Ana Salgado, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, and Sia Kolkovska. 2025. Parallel sense-annotated corpus ELEXIS-WSD 1.3. Slovenian language resource repository CLARIN.SI.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, 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*, pages 377–395.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Ranka Stanković, Cvetana Krstev, Saša Petalinkar, Milica Ikonić Nešić, Aleksandra Marković, Marina Bagi, Marijana Đukić, and Jelena Bogdanović. 2026. SrELEXIS-WSD: Hybrid Semi-Automated WSD for Serbian with Large Language Models—Results and Challenges. In *4th UniDive General Meeting*.