Linking MWE occurrences in corpora with their sense inventory entries using Linguistic Linked Data technology

Ranka Stanković[1], **Verginica Barbu Mititelu**[3], **Jan Odijk**[4], **Voula Giouli**[5], **Carole Tiberius**[6], **Milica Ikonić Nešić**[1,2]

[1]University of Belgrade and Jerteh - Language Resources and Technologies Society, Serbia;
[3]RACAI, Bucharest, Romania; [4]Utrecht University, Netherlands;
[5]Aristotle University of Thessaloniki and ILSP, ATHENA RC, Greece;
[6]Dutch Language Institute/Leiden University Centre for Linguistics, The Netherlands.

Relevant UniDive working groups: WG2

Introduction

A persistent challenge in computational lexicography lies in bridging lexical resources and corpus data. While MWEs are frequently identified and analyzed in corpora, lexical resources often lack explicit links to corpus attestations and contextual evidence (Mititelu et al., 2025). Conversely, corpora contain numerous occurrences of MWEs that are not systematically connected to lexical entries. Building on the work done in Task T2.3.3 of the UniDive COST Action (Savary et al., 2024), this work investigates how Linguistic Linked Open Data (LLOD) technologies can support standardized representations of MWEs, how they can be systematically linked with each other, to real corpus usage, to knowledge graphs (KGs) and how such infrastructures support integrated querying and exploration across lexicons and corpora.

Our focus here is on cross-lingual analysis of interlinking aligned corpora annotated with MWEs by using community LLD standards such as the NLP Interchange Format (Hellmann et al., 2013, NIF), designed for web-based linguistic annotation, and CoNLL-RDF (Chiarcos and Glaser, 2020), a vocabulary tailored for compatibility with tabular NLP formats, including CoNLL, Universal Dependencies (CoNLL-U), and PARSEME (Parseme-TSV). For modeling MWEs we focus on OntoLex,[1] a widely used community standard for lexical data available in RDF (McCrae et al., 2017).

For this Proof-of-concept lexicon encoding of MWEs and lexicon-corpus interface, we use the ELEXIS-WSD dataset[2] (from task T2.2 in WG2).

[1]https://www.w3.org/2016/05/ontolex
[2]http://hdl.handle.net/11356/2101

1 Methodology

The modeling framework is based on OntoLex-Lemon, where MWEs are represented as instances of `ontolex:MultiwordExpression`. Their internal structure is modeled using the `decomp` module, which specifies constituent components and their relation to the lexical entry (Fig. 1). Cross-lingual relations are encoded using the `vartrans` module, linking lexical senses across languages (Bosque-Gil et al., 2019).

Corpus evidence is integrated using the OntoLex-FrAC vocabulary (Chiarcos et al., 2022), which allows representation of attestations and frequency information. Corpus annotations are encoded in the NLP Interchange Format (NIF), enabling precise linking between lexical entries and their occurrences in text via character offsets and context identifiers.

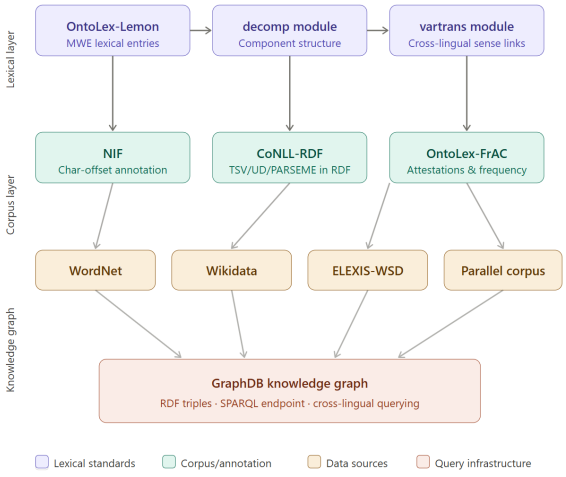

Figure 1: An overview of modeling framework

A key extension presented in this work is the deployment of the integrated data in a knowledge graph environment using GraphDB. Lexical entries, corpus annotations, and semantic links

are stored as RDF triples and exposed through a SPARQL endpoint.[3] This enables complex queries over MWEs, including the retrieval of corpus attestations, exploration of translation equivalents, and navigation across linked lexical and corpus resources.

2 MWEs as LLOD

The approach is illustrated using resources developed in Task 2.2 of UniDive (Čibej et al., 2026), besides sense inventories and corpora annotated with multiple layers, including MWEs and named entities linked to the knowledge base WikiData.

Lexical entries are enriched with canonical forms, part-of-speech information, semantic references (e.g. WordNet synsets and Wikidata identifiers), and decomposition into components. For example, the English expression *world war* is represented as a multiword lexical entry linked to its components and to corresponding expressions in other languages, such as Serbian *svetski rat*, or Romanian *război mondial*. Since senses for these expressions are annotated with the WordNet interlingual index `ENG30-00996817-n`, the sense translation link can be established.

```
:le_world_war-en
 a ontolex:LexicalEntry, ontolex:MultiwordExpression;
 ontolex:canonicalForm
  [ontolex:writtenRep "world war"@en];
 lexinfo:partOfSpeech lexinfo:noun;
 ontolex:sense :se_world_war-en;
 decomp:constituent :cm_world-en;
 decomp:constituent :cm_war-en;
 rdf:_1 :le_world-en; # lexical
 rdf:_2 :le_war-en.  # entries

:se_world_war-en a ontolex:LexicalSense ;
  ontolex:isSenseOf :le_world_war-en ;
  ontolex:reference :wn30_eng30_02202047_n .

:wn30_eng30_02202047_n a ontolex:LexicalConcept ;
 dcterms:identifier "ENG30-02202047-n" ;
 rdfs:label "world war"@en .

# component of cannonical form
:cm_world-en  a decomp:Component;
 decomp:correspondsTo :le_world-en.
  ...
:le_svetski_rat-sr a [...];
 ontolex:canonicalForm
  [ontolex:writtenRep "svetski rat"@sr];

:le_razboi_mondial-ro a [...];
 ontolex:canonicalForm
 [ontolex:writtenRep "război mondial"@ro];
```

The following example models a translation set that groups cross-lingual sense correspondences, where English and Romanian lexical senses are linked through individual `vartrans:Translation` relations to the corresponding Serbian sense.

```
:tranSetEN-SR a vartrans:TranslationSet ;
 dc:source <http://elexis-wsd-ds-url>;
 vartrans:trans
  :se_world_war-en-se_svetski_rat-sr-trans,
  :se_razboi_mondial-ro-se_svetski_rat-sr-trans .

:se_world_war-en-se_svetski_rat-sr-trans
 a vartrans:Translation ;
 vartrans:source :se_world_war-en ;
 vartrans:target :se_svetski_rat-sr .

:se_razboi_mondial-ro-se_svetski_rat-sr-trans
 a vartrans:Translation ;
 vartrans:source :se_razboi_mondial-ro ;
 vartrans:target :se_svetski_rat-sr .
```

One option of linking is the OntoLex-FrAC vocabulary, which implements the lexicon-corpus interface, e.g., a corpus example for sentence `12.en`: *In 1914 the First World War broke out.*, and corresponding translation equivalents can be encoded as follows:

```
:le_svetski_rat frac:attestation [frac:quotation
 "In 1914 the First World War broke out."@en;
 frac:observedIn :EWSD].
:le_svetski_rat-sr frac:attestation [frac:quotation
 "Prvi svetski rat izbio je 1914."@sr; ...].
:le_razboi_mondial-ro frac:attestation
 [frac:quotation "În 1914 a izbucnit Primul Război
 Mondial."@ro; ...].
:le_uitbreken_wereldoorlog-nl frac:attestation
 [frac:quotation "In 1914 brak de Eerste
 Wereldoorlog uit."@nl; ...].
```

The second option is to represent sentences using NIF and the CoNLL-RDF, allowing direct linking between lexical entries and specific occurrences in text (Section 3).

3 Parallel corpus as LLOD

Using off-the-shelf RDF technologies and the OntoLex-FrAC vocabulary, corpus data and sense inventories can be integrated into a unified, queryable representation. On the corpus side, NIF (Hellmann et al., 2013; Brümmer, 2015; Cimiano et al., 2020) facilitates the incorporation of linguistic annotations into the LLOD Cloud, ensuring accessibility and reusability of language resources across diverse applications and domains.

A minimal subset of NIF constitutes the basis of CoNLL-RDF, which offers a standardized format for representing syntactic and morphological annotations, fostering consistency and compatibility in linguistic data representation. Using the CoNLL-RDF (Chiarcos and Fäth, 2017) data can be generated on the fly from CoNLL corpora, and such data can be serialized back to the original TSV formats. The RDF representation of TSV data provides a flexible and powerful technique for querying and consuming such data, and for querying and integrating it with the associated lexical resources.

[3] Apart from SPARQL endpoint `http://graphdb.jerteh.rs/sparql?repositoryId=elexis-wsd-nif`, the NIF+Ontolex version of ELEXIS-WSD will be published at CLARIN node.

Through practical implementations and experiments, we demonstrate the effectiveness of incorporating LLOD principles, NIF, and CoNLL-RDF in aligned parallel corpora annotation. The proposed approach enhances interoperability and accessibility of linguistic resources but also lays the foundation for more comprehensive and nuanced language technologies. The following example illustrates the annotated MWE spans *First World War* linked to the Wikidata item `Q361`, but the same item is also linked in `12.sr` *Prvi svetski rat*, in `12.ro` *Primul Război Mondial*, and in `12.nl` *Eerste Wereldoorlog*. The offsets in example correspond to character positions.

```
<http://url> a nif:ContextCollection ;
 nif:hasContext <http://url/enwsd>  .
<http://url/enwsd> a nif:Context,
        nif:OffsetBasedString ;
 nif:beginIndex "0"  ;
 nif:endIndex "38"  ;
 nif:isString "In 1914 the First World War
  broke out." .
[...]
<http://url/enwsd#offset_18_27_0> a
 nif:OffsetBasedString, nif:Phrase ;
 nif:anchorOf "First World War"@en ;
 nif:beginIndex "18";
 nif:endIndex "27" ;
 nif:referenceContext <http://url/enwsd> ;
 nif:taMsClassRef/itsrdf:taIdentRef wd:Q361;
 itsrdf:taClassRef dbo:Event,wd:Q1190554,
   <http://nerd.eurecom.fr/ontology#Event>.
```

The findings presented in this paper contribute to the ongoing discourse on leveraging LLOD in the realm of aligned parallel corpora with WSD annotation, paving the way for more robust and efficient NLP applications. By using NIF, OLiA, OntoLex, CoNLL-RDF and inspired by RuSemCor,[4] an open, manually crafted WSD Russian corpus that aligns tokens from the corpus with senses defined in the Russian WordNet (Kirillovich et al., 2025). The resource is formally represented using Semantic Web standards, including the NIF, OLiA, OntoLex, and Global WordNet ontologies, and is integrated into LLOD cloud. The property `conll:LEXICAL_SENSE_URI` is used for establishing link between text span (or token) from corpus and entry in the sense inventory.

```
<http://url/enwsd#offset_24_27_0> a
nif:OffsetBasedString, nif:Phrase ;
nif:anchorOf "world war" ;
nif:beginIndex "24";
nif:endIndex "27" ;
nif:referenceContext <http://url/enwsd> ;
...
conll:LEXICAL_SENSE_URI :se_world_war-en.
```

The integration of these resources in GraphDB enables querying such links through SPARQL, supporting retrieval of MWEs across corpora, languages, and semantic layers. The queryability will be demonstrated with set of examples available in T2.3.3 page.[5]

In lexicographic practice, the resource can support the creation, enrichment, and maintenance of digital dictionaries by linking lexical entries with corpus attestations, semantic relations, translations, and contextual usage examples in a machine-readable format. Through RDF-based interoperability, lexicographers can efficiently connect lexical data with external resources such as WordNet, or Wikidata, facilitating semantic exploration and multilingual alignment. In downstream NLP tasks, the resource can serve as structured semantic knowledge for applications such as named entity recognition, relation extraction, WSD, semantic search, question answering, and information retrieval. The use of NIF allows direct alignment between textual spans and linguistic annotations, while OntoLex provides a formal representation of lexical meanings and lexical-semantic relations.

4 Conclusion

This work contributes to the development of interoperable lexicographic infrastructures by combining lexical modeling, corpus annotation, and knowledge graph technologies. First, it demonstrates how MWEs can be represented as structured lexical entities in OntoLex-Lemon, while preserving their internal structure and semantic interpretation. Second, it strengthens the lexicon–corpus interface by linking lexical entries with corpus attestations encoded in NIF. Third, it introduces a knowledge graph–based approach that enables advanced querying and exploration of MWEs using SPARQL in GraphDB. Chiarcos and Schenk (2018) demonstrated the usefulness and practicability of ACoLi CoNLL libraries on the Universal Dependency corpus, providing a path for our further research.

By integrating lexical resources, corpora, and external knowledge bases such as Wikidata, the proposed framework supports multilingual phraseological research and facilitates the discovery, analysis, and reuse of MWEs across languages and datasets. This approach operationalizes the lexicon–corpus interface within the LLOD paradigm by enabling bidirectional navigation between textual evidence and structured lexical-semantic representations.

[4] https://github.com/LLOD-Ru/rusemcor

[5] https://unidive.lisn.upsaclay.fr/doku.php?id=wg2:t233

Acknowledgements

This research was supported by the COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive); Science Fund of the Republic of Serbia, #GRANT 7276, Text Embeddings- Serbian Language Applications- TESLA. We would like to express our sincere gratitude to Jaka Čibej for the considerable effort and dedication invested in preparing the ELEXIS-WSD version for publication.

References

Julia Bosque-Gil, Dorielle Lonke, Ilan Kernerman, and Jorge Gracia. 2019. Validating the Ontolex-lemon Lexicography Module with K Dictionaries' Multilingual Data. In *Proceedings of Electronic Lexicography in the 21st Century Conference 2019*, ART-2019-123124, pages 726–746.

Martin Brümmer. 2015. Expanding The NIF Ecosystem-Corpus Conversion, Parsing And Processing Using The NLP Interchange Format 2.0.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Christian Chiarcos and Luis Glaser. 2020. A Tree Extension for CoNLL-RDF. In *Proceedings of the 12th LREC*, pages 7161–7169.

Christian Chiarcos and Niko Schenk. 2018. The ACoLi CoNLL libraries: Beyond Tab-separated Values. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Linked Data-Based NLP Workflows. *Linguistic Linked Data: Representation, Generation and Applications*, pages 197–211.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

Alexander Kirillovich, Ilia Karpov, Natalia Loukachevitch, Maksim Kulaev, and Dmitry Ilvovsky. 2025. Rusemcor: A word sense disambiguation corpus for russian. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6438–6443, New York, NY, USA. Association for Computing Machinery.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Verginica Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic, and Ivelina Stoyanova. 2025. Survey on lexical resources focused on multiword expressions for the purposes of NLP. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 41–57, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

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.

Jaka Čibej, Simon Krek, Carole Tiberius, Irina Lobzhanidze, Verginica Barbu Mititelu, Jelena Kallas, Kertu Saul, Kadri Muischnek, Ranka Stanković, Cvetana Krstev, Aleksandra Marković, Ana Ostroški Anić, Bartłomiej Alberski, Vladimir Cvetkoski, Voula Giouli, Rusudan Makhachashvili, and Olha Kanichsheva. 2026. Extension of the ELEXIS-WSD Parallel Sense-Annotated Corpus Within UniDive: New Languages and Layers. In *UniDive 4th Workshop, Bucharest*.