Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora

Ranka Stanković
University of Belgrade
F.of Mining and Geology
Belgrade, Serbia
ranka@rgf.rs

Christian Chiarcos
University of Augsburg
App.Computational Linguistics
Augsburg, Germany
christian.chiarcos@uni-a.de

Milica Ikonić Nešić
University of Belgrade
Faculty of Philology
Belgrade, Serbia
milica@jerteh.rs

Relevant UniDive working groups: WG2

1 Introduction

In this abstract, we elaborate use of Linked Data for the lexicon-corpus interface, aiming at interlinking MWE lexicon entries with their occurrences in corpora. The cross-lingual analysis of idiosyncratic constructions can be supported by publishing aligned and annotated corpus data as Linked Data employing community standards such as the NLP Interchange Format (Hellmann et al., 2013, NIF) and CoNLL-RDF, a minimal NIF subset designed for compatibility with tab-separated formats used in NLP ("CoNLL"), Universal Dependencies ("CoNLL-U") and Parseme ("Parseme-TSV").

The goal of the UniDive T2.2 *Design of a lexicon-corpus interface* is to extend the ELEXIS-WSD Parallel Sense-Annotated Corpus (Federico et al., 2021), available in its current version (1.1) at the CLARIN.SI repository: http://hdl.handle.net/11356/1842. The corpus consists of 2,024 sentences for 10 languages and a sense repository. New languages should be included, but also the corpus annotation will be upgraded to allow for interlinking MWE lexicon entries with their occurrences in corpora. Additional annotation layers are envisaged in the context of UniDive. Moreover, these resources should be also published as Linked Data (using NIF, Sect. 3) to facilitate its linking with the sense repository of the corpus (using OntoLex, Sect. 2).

2 MWE dictionaries as Linked Data

Linked Data constitutes a suitable common representation framework for various types of dictionaries that have been built by following different perspectives while retaining all the benefits related to interoperability, visibility, and NLP-services compliance. The seamless integration with other internal and external resources (via links among entities, expressed for example in RDF) allows for a natural graph-based representation of dictionary data based on Web standards (Gracia et al., 2017).

For modeling the ELEXIS sense inventory in RDF, we focus on OntoLex,[1] a widely used community standard for machine-readable lexical resources in the context of RDF, Linked Data, and Semantic Web technologies (McCrae et al., 2017) and the basis of most lexical data available in RDF.[2]

OntoLex modules relevant to MWEs include the core module **Lemon** (general data structures), the module **Decomp** (for the internal structure and combinatory semantics of MWEs), and the **Morph** module (MWE morphology), **Lexicog** module for lexicography (Bosque-Gil et al., 2019). Furthermore, Ontolex is extended with a new module for Frequency, Attestations, and Corpus-based Information (**FrAC**) (Chiarcos et al., 2022a,c),[3] which adds support for linking lexicons with corpora, collocations analysis, enabling RDF-based web services to exchange corpus queries and response data dynamically, and other aspects of information relevant to the joint work with corpora and dictionaries.

Taking the BPMLOD Guidelines for Linguistic Linked Data Generation https://bpmlod.github.io/Bilingual_Dictionaries_Report/ as a starting point for modelling the multilingual sense repository (MSR) to be developed within Task 2.2, the following example gives an idea of how the lexical entry for MWE *blood pressure* can be

[1] https://www.w3.org/2016/05/ontolex

[2] Another candidate vocabulary is the Data Model for Lexicography (DMLex), developed under participation of ELEXIS contributors, which, however, is not formally published yet, and which also defines a mapping to OntoLex.

[3] The current draft version of the FrAC specification is found under https://github.com/ontolex/frequency-attestation-corpus-information/

represented in RDF along with translations and links between senses and DBpedia entries.

```
:le_blood_pressure
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "blood pressure"@en];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference <https://
    dbpedia.org/page/Blood_pressure>];
  decomp:constituent :cm_blood;
  decomp:constituent :cm_pressure;
  rdf:_1 :le_blood; # lexical
  rdf:_2 :le_pressure.  # entries

# component of cannonical form
:cm_blood  a decomp:Component;
  decomp:correspondsTo :le_blood.
  ...
:le_krvni_tlak a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "krvni tlak"@sl];
  ...
:tranSetEN-SL a vartrans:TranslationSet ;
 dc:source
   <http://hdl.handle.net/11356/1842> ;
  ...
# simplified naming
:tranSetEN-SL vartrans:trans
 blood_pressure-ensns-krvni_tlak-slsns .
:blood_pressure-ensns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_blood_pressure .
:krvni_tlak-slsns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_krvni_tlak .
:blood_pressure-ensns-krvni_tlak-sns-trans
 a vartrans:Translation ;
 vartrans:source :blood_pressure-ensns ;
 vartrans:target :krvni_tlak-slsns .
```

For such data, the OntoLex-FrAC vocabulary implements the lexicon-corpus interface, e.g., a corpus example for sentence *1573 "Physical examination may sometimes reveal low blood pressure, high heart rate, or low oxygen saturation."* (English), resp., *"Telesni pregled včasih razkrije nizek krvni tlak, visok srčni utrip ali nizko nasičenost s kisikom. "* (Slovenian) can be encoded as follows:

```
# multiword inflected form attestation
:le_blood_pressure
 frac:attestation [
 frac:quotation "Physical examination
 may sometimes reveal low blood pressure,
 high heart rate, or low oxygen
 saturation."@en;
 frac:observedIn :EWSD].

:le_krvni_tlak
 frac:attestation [
 frac:quotation "Telesni pregled včasih
 razkrije nizek krvni tlak, visok srčni
 utrip ali nizko nasičenost s kisikom."@sl;
 frac:observedIn :EWSD].
```

Beyond that, OntoLex-FrAC can also be used for frequencies (Chiarcos et al., 2022a), collocations (Chiarcos et al., 2022b), and embeddings (Chiarcos et al., 2021). Along with the poster, illustrative examples will be made available online for the presentation.

3 Parallel corpus as Linked Data

The motivation for publishing parallel corpora as Linked Data lies in the benefits of increased accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science. These motivations collectively contribute to advancing linguistic research, language technology, and cross-disciplinary insights. In the rapidly evolving landscape of language technologies, the integration of Linked Data has emerged as a pivotal force, revolutionizing the way linguistic resources are managed and utilized. This paper explores the linked data version and the extension of the ELEXIS-WSD Parallel Sense-Annotated Corpus being developed within Task2.2 of WG2 using NIF and the CoNLL-RDF. Using off-the-shelf RDF technologies and the OntoLex-FrAC vocabulary, these data can be linked and queried and processed together.

On the corpus side, NIF (Hellmann et al., 2012; Brümmer, 2015; Cimiano et al., 2020) facilitates the incorporation of linguistic annotations into the Linked Data 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 library (Chiarcos and Fäth, 2017), CoNLL-RDF (and thus, NIF) data can be generated on the fly from CoNLL corpora, and such data can be serialized back to the original TSV formats. While TSV data is optimal for storing and exchanging such data, its RDF representation provides a flexible and powerful technique for querying and consuming such data, and in particular, for querying and integrating it with the associated lexical resources. Through practical implementations and experiments, we demonstrate the efficiency of incorporating Linked Data principles, NIF, and CoNLL-RDF in aligned parallel corpora annotation. The proposed approach not only enhances the interoperability and accessibility of linguistic resources but also lays the foundation for more comprehensive and nuanced language technologies. The findings presented in this paper contribute to the ongoing discourse on leveraging Linked Data in the realm of aligned parallel corpora annotation, paving the way for more robust and efficient NLP applications.

```
<http://url> a nif:ContextCollection ;
    nif:hasContext <http://url/enwsd>  .

<http://url/enwsd> a nif:Context,
        nif:OffsetBasedString ;
    nif:beginIndex "0"  ;
    nif:endIndex "49"  ;
    nif:isString "He is named after
    the astronomer Galileo Galilei." .

<http://url/enwsd#offset_0_49_0> a
 nif:OffsetBasedString, nif:Phrase ;
 nif:anchorOf "Galileo Galilei." ;
 nif:beginIndex "33";
 nif:endIndex "49" ;
 nif:referenceContext <http://url/enwsd> ;
 nif:taMsClassRef/itsrdf:taIdentRef wd:Q307;
 itsrdf:taClassRef dbo:Person,wd:Q5,
   <http://nerd.eurecom.fr/ontology#Person>.
```

For establishing links between sentences that are translation equivalents, *skos:closeMatch* from *SKOS* (Simple Knowledge Organization System)[4] can be used: *skos:closeMatch* indicates that two objects are sufficiently similar that they can be used alternately in applications.

Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia, #GRANT 7276, Text Embeddings - Serbian Language Applications - TESLA, and COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

References

Julia Bosque-Gil, Dorielle Lonke, I Kernerman, and J Gracia. 2019. Validating the ontolex-lemon lexicography module with k dictionaries"multilingual data. In *Electron. lexicogr. 21st cent., Proc. eLex conf.*, ART-2019-123124.

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. 2022a. 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, Thierry Declerck, and Maxim Ionov. 2021. Embeddings for the lexicon: Modelling and representation. In *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*, pages 13–19.

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, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Anas Fahad Khan, and Ciprian-Octavian Truica. 2022b. Modelling collocations in OntoLex-FrAC. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 10.

Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022c. Modelling Collocations in OntoLex-FrAC. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18.

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.

Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Proceedings of the eLex 2021 conference*, pages 377–395. Lexical Computing.

Jorge Gracia, Ilan Kernerman, and Julia Bosque-Gil. 2017. Toward linked data-native dictionaries. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

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.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. Nif combinator: Combining nlp tool output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

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.

[4] https://www.w3.org/TR/skos-reference/