WG2

UniDive

2nd General Meeting

University of Naples "L'Orientale

Naples, Italy, 8-9 February 2024

https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

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

Ranka Stanković[1], Christian Chiarcos[2], Milica Ikonić Nešić[1]

[1]University of Belgrade, Serbia; [2]University of Augsburg, Germany

Introduction

- Linked Data (LD) for the lexicon-corpus interface: interlinking MWE lexicon entries with their occurrences in corpora
- Publishing of aligned and annotated corpus as LD employing NLP Interchange Format (NIF) and CoNLL-RDF
- UniDive T2.2: Design of a lexicon-corpus interface is to extend the ELEXIS-WSD Parallel Sense-Annotated Corpus (2,024 sentences for 10 languages and a sense repository.
- Apart from new languages and new annotation layers, also corpus annotation shlould be upgraded to allow for interlinking MWE lexicon entries with their occurrences in corpora (as LD).

MWE dictionaries as Linked Data

- Modeling: OntoLex, https://www.w3.org/2016/05/ontolex (possible alternative Data Model for Lexicography (DMLex), v1.0 draft)
- Core module Lemon (general data structures),
- Decomp (for the internal structure and combinatory semantics of MWEs),
- Morph module (MWE morphology),
- Lexicog module for lexicography,
- FRaC (Frequency, Attestations, and Corpus-based Information

- MWE lexical entry example: **blood pressure** with translations and links between senses and DBpedia entries.

Ontolex-lemon

```
: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/...> ;
  …
# 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 .
```

Decomp

Vartrans

Parallel corpus as Linked Data

- Parallel corpora as LD benefits of increased accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science.
- Through practical implementations and experiments, we demonstrate the efficiency of incorporating LD principles, NIF, and CoNLL-RDF in aligned parallel corpora annotation.
- The findings presented in this paper contribute to the ongoing discourse on leveraging LD 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) can be used (indicates that two objects are sufficiently similar that they can be used alternately in applications).

```
# 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].
```

Ontolex-FRaC

This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications - TESLA.

Tesla

Science Fund of the Republic of Serbia

Nexus Linguarum

ranka@rgf.rs

christian_chiarcos@uni-a.de

milica.ikonic.nesic@fil.bg.ac.rs

Universität Augsburg University