cography
tain
ions
Speech
ologisms
Idioms
NLP
Etymology
Glossary
Lemma
gin
Corpora
Dictionary
Word
Meaning
Lexicon
ble
Pronounciation
Headword
Definition
amples
Entry
exicology
Dictionary Use
Lexical Resources

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Interlinking Slovene Language Datasets

Bajčetić L.[1], **Declerck T.**[1,2]

[1]Austrian Centrefor Digital Humanities and Cultural Heritage.Austrian Academy of Sciences,Austria
[2] DFKI GmbH, Multilinguality and Language TechnologyLab, Germany

Abstract

We present the current implementation state of our work consisting in interlinking language data and linguistic information included in different types of Slovenian language resources. The types of resources we currently deal with are a lexical database (which also contains collocations and example sentences), a morphological lexicon, and the Slovene WordNet. We first transform the encoding of the original data into the OntoLex-Lemon model and map the different descriptors used in the original sources onto the LexInfo vocabulary. This harmonization step is enabling the interlinking of the various types of information included in the different resources, by using relations defined in OntoLex-Lemon. As a result, we obtain a partial merging of the information that was originally distributed over different resources, which is leading to a cross-enrichment of those original data sources. A final goal of the presented work is to publish the linked and merged Slovene linguistic datasets in the Linguistic Linked Open Data cloud.

Keywords: Slovenian Language Data; interlinking; OntoLex-Lemon; LexInfo

1 Introduction

In the context of approaches aiming at the generation of a densely linked dataset of language resources, we are considering different types of Slovenian language resources, consisting of lexical, morphological, and conceptual data. The Slovene data sets we are including in our current work are:

1. Slovene Lexical Database (SLD) - a lexical database, which deals with collocations (Gantar & Krek 2011)
2. Sloleks - a morphological lexicon (Dobrovoljc et al. 2017)
3. sloWNet - Slovenian WordNet (Fišer et al. 2012)

Linking and merging processes have been implemented for those three resources, and we are planning to extend the work to other (types of) resources. For the cross-linking and merging of the different types of language data we are making use of the OntoLex-Lemon framework (Cimiano et al. 2016). For the present work, we focus only on nouns. The use of OntoLex-Lemon for representing lexicographic data has been previously presented and discussed in (Declerck et. 2017; Tiberius & Declerck 2017). The relevance of OntoLex-Lemon for the representation of WordNet data and for interlinking and merging WordNet and lexicographic data for Romance languages has been demonstrated in (Racioppa & Declerck 2019) and we anticipated similar results for the Slovenian language.

In the following sections we present first the selected Slovenian resources. We continue with a brief description of the OntoLex-Lemon model and the LexInfo ontological vocabulary, which is providing data categories for the model. We then present some results of the linking and merging processes, as they are represented in OntoLex-Lemon. We close the paper with a description of the planned next steps.

2 The selected Slovene Datasets

In the following sections we briefly present the Slovenian data sets we are currently dealing with, and which are representing a wide coverage of different types of linguistic information. For all those datasets we give as an example the way they encode information related to the word "alergija" (*allergy*).

2.1 The Slovene Lexical Database (SLD)

The Slovene Lexical Database (SLD) is a lexical-conceptual resource, structured as a network of interrelated semantic and syntactic information about a word. SLD contains dictionary-type of information on words and word combinations, for example senses, collocations, example sentences, syntactic patterns, grammatical information, etc.

The database was compiled from the Gigafida corpus (Logar & Kosem 2011), a recent generation of Slovene corpora which contains 1,134,693,933 words from 38,310 texts of different genres, including Internet content. The core element of the resource is the lexical unit which includes all senses of the headword, multi-word expressions and phraseological units. SLD contains two types of information which are designed for two types of users: the first is the lexico-grammatical information that is intended for human users and comes in the form of sense descriptions as well as collocations and typical examples from the corpus, which are both attributed to particular senses and syntactic patterns of the lemma. The second type of information is devised for natural language processing tools. It includes the formal encoding of syntactic patterns at the clause and phrasal level (syntactic structures) as well as the formal encoding of

semantic arguments and their types. For our purpose, we have extracted all the lemmas with collocations and example sentences stored alongside the syntactic pattern which they exhibit. This way our integrated resource can be enriched with a multitude of example sentences which can potentially be used to improve disambiguation.
The original encoding for the word *alergija* in SLD is given in Table 1 just below:

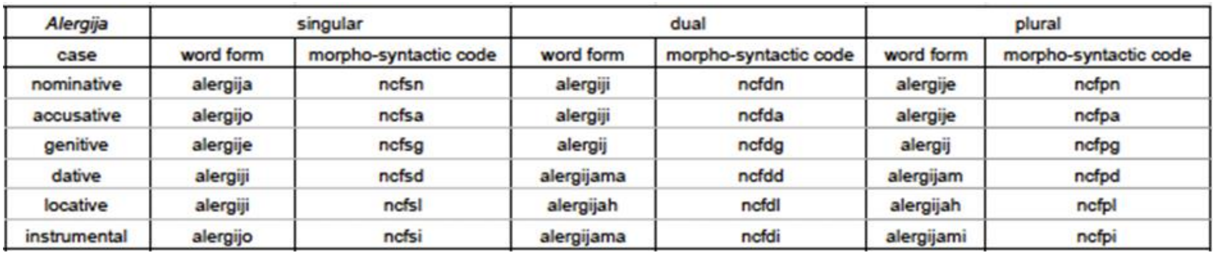
```
<zapis>alergija</zapis>
<iztocnica>alergija</iztocnica>
</oblika>
<zaglavje>
<besvrs>samostalnik</besvrs>
</zaglavje>
</glava>
<geslo>
<pomen>
<indikator>preobčutljivost organizma</indikator><oznaka tip="podrocje">zdravje</oznaka>
<pomenska_shema/><definicija1>zdravstveno stanje, ki se kaže kot preobčutljivost organizma na določeno snov ali hrano, s katero
pride v stik</definicija1>
<skladenjske_skupine>
<skladenjska_struktura>
<struktura>sbz0 SBZ2</struktura>
<kolokacije>
<kolokacija><k>zdravljenje, simptom, znak, ugotavljanje, diagnoza, odkrivanje</k> alergije</kolokacija>
<kolokacija><k>povzročitelj, sprožilec</k> alergije</kolokacija>
<kolokacija><k>nastanek, pojavljanje, pojav, porast, izbruh</k> alergije</kolokacija>
<kolokacija><k>razvoj, posledica, preprečevanje, vzrok, pogostnost, preprečitev</k> alergije</kolokacija>
<kolokacija><k>oblika, vrsta, primer, tip</k> alergije</kolokacija>
<kolokacija><k>nevarnost, napad, čas</k> alergije</kolokacija>
</kolokacije>
<zgledi>
```

Table 1: The entry for the word *alergija* in the Slovene Lexical Database

As the reader can observe, the used descriptors are in Slovene. There is a need to map those descriptors (or tags) to an interoperable vocabulary, which in our work is given by LexInfo (see Section **Fehler! Verweisquelle konnte nicht gefunden werden.**.2). This mapping onto LexInfo is particularly relevant for the descriptor used within the "<struktura>" tag for marking the syntactic structure of a collocation, which is represented in SLD as an abstract pattern. See (Fišer et al. 2012) for a more in-depth explanation of the Slovene Lexical Database, where the authors already present an approach for relating SLD and sloWNet.

2.2 The Slovene Morphological Lexicon Sloleks

Sloleks is a large open-source machine-readable morphological lexicon for the Slovene language (Dobrovoljc et al. 2008). The lexicon provides inflectional, derivational, and grammatical information, formally represented within the XML serialization of the standardized LMF framework, which is described in (Francopoulo et al. 2006). For morphologically rich languages, such as Slovene, describing morphological paradigms of inflected parts of speech is a crucial step in creating the language model. Therefore, databases such as Sloleks are incredibly valuable. Sloleks is designed as a digital resource for computational linguistics, which means it contains a huge collection of systematically described morphological patterns in machine-readable format, resulting in almost 2 800 000 inflected forms. An entry includes the basic form (the lemma) of the word, its inflected forms (the inflectional paradigm) and related morphological information. Since we only focus on nouns for now, we have extracted all the noun entries which contain the lemma and all the inflected forms marked with case and number. Other information we have extracted is gender and pronunciation information, which can hopefully be used in the future, as well as some information of corpus frequency, also included in Sloleks. Table 2 just below displays the (simplified) Sloleks encoding of "alergija", in a tabular format we designed for the purpose of this paper.



Alergija	singular		dual		plural	
case	word form	morpho-syntactic code	word form	morpho-syntactic code	word form	morpho-syntactic code
nominative	alergija	ncfsn	alergiji	ncfdn	alergije	ncfpn
accusative	alergijo	ncfsa	alergiji	ncfda	alergije	ncfpa
genitive	alergije	ncfsg	alergij	ncfdg	alergij	ncfpg
dative	alergiji	ncfsd	alergijama	ncfdd	alergijam	ncfpd
locative	alergiji	ncfsl	alergijah	ncfdl	alergijah	ncfpl
instrumental	alergijo	ncfsi	alergijama	ncfdi	alergijami	ncfpi

Table 2: The morphological variants of the entry "alergija" in Sloleks, with their original abbreviated encoding, which has been mapped onto LexInfo (see Section 3.2, Example 1)

2.3 The Semantic Lexicon of Slovene: sloWNet

sloWNet is the Slovene WordNet, developed in the expand approach: it contains the complete Princeton WordNet 3.0 and over 70,000 Slovene literals. These literals have been added automatically using different types of existing resources, such as bilingual dictionaries, parallel corpora, and Wikipedia. For the scope of this work we have extracted only the Slovene literals with their synset ids. For future work it would be interesting to see to what extent the semantic information encoded in PWN can be used for Slovene. How the lemma *alergija* is represented in sloWNet is shown in Table 3 just below:

```
<LexicalEntry id ='w1167167'>
  <Lemma writtenForm='alergija' partOfSpeech='n'/>
  <Sense id='w1167167_05653475-n' synset='slv-10-05653475-n'/>
  <Sense id='w1167167_14532816-n' synset='slv-10-14532816-n'/>
  <Sense id='w1167167_14533796-n' synset='slv-10-14533796-n'/>
</LexicalEntry>
```

Table 3: The encoding of the word *alergija* in sloWNet, in the XML serialization of the LMF model

In Table 3, we can see that the linguistic information is poor in this resource. Only the part-of-speech is indicated for the corresponding lemma. Our work consists in linking the sloWNet concepts to the full lexical description available in Sloleks. This can be straightforwardly done, once both resources have been transformed onto the OntoLex-Lemon model.

3 OntoLex-Lemon and LexInfo

We present briefly the instruments used for transforming the original Slovene datasets into a harmonized shared representation. On the one hand we have the OntoLex-Lemon model, which we deploy for representing the lexical and conceptual data, and on the other hand the LexInfo vocabulary, which was, among others, developed for providing a set of data categories for use in OntoLex-Lemon.

3.1 Ontolex-Lemon

OntoLex-Lemon is a further development of the "Lexicon Model for Ontologies" (*lemon*, see McCrae et al. 2012). Both *lemon* and the OntoLex-Lemon model, which is resulting from a W3C Community Group, were originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description (Cimiano et al. 2016). This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies. The main organizing unit for those linguistic descriptions is the LexicalEntry class, which enables the representation of morphological patterns for each entry (a multiword expression, a word, or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the ontolex:denotes property or is mediated by the LexicalSense or the LexicalConcept classes, as this is represented in Figure 1, which displays the core module of the model. A major difference between *lemon* and OntoLex-Lemon is that the latter includes an explicit way to encode conceptual hierarchies, using the SKOS[1] standard. As can be seen in Figure 1, lexical entries can be linked via the ontolex:evokes property to such SKOS concepts, which we use to represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the ontolex:reference property or mediated by the instances of the ontolex:LexicalSense class.

OntoLex-Lemon comes with additional modules, one of them being used in our transformation and harmonization exercise: the decomp module. This module supports the representation of elements of a multi-word or compound lexical entry. As can be seen in Figure 2 below, the decomp module makes use of the decomp:subterm property in order to indicate that a multi-word lexical entry contains other entries. The decomp: Component class is there for collecting the components of the lexical entry as the individual tokens (particular realizations of lexical entries) that compose that compound lexical entry. This module is relevant for representing the collocations that are included in SLD, as one view on collocations can be that they are a type of Multi Word Expressions (MWE).

[1] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/) [31.07.2020]

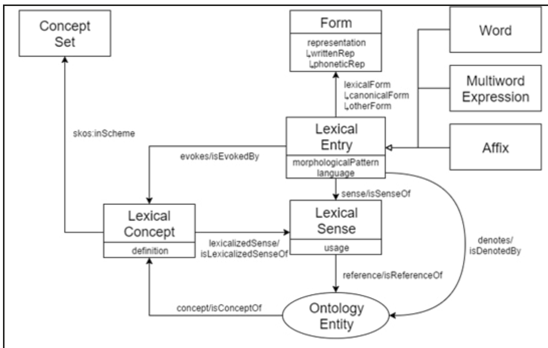


Figure 1: The core module of OntoLex-Lemon (taken from https://www.w3.org/2016/05/ontolex/)

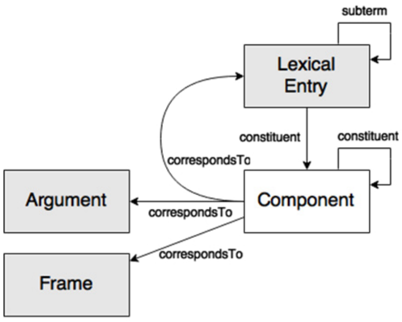


Figure 2: The decomp module of OntoLex-Lemon (taken from https://www.w3.org/2016/05/ontolex/)

3.2 LexInfo

As already stated, LexInfo is an ontology that was defined to provide data categories for the *lemon* model, and which has been updated with the new OntoLex-Lemon model of the W3C Ontolex community group. LexInfo is designed as an ontology and is written using the same W3C standards as OntoLex-Lemon, being RDF, RDF(s) or OWL, making thus use of resource-describing graphs as the basic representation instrument. We deploy LexInfo to harmonize all the descriptors (tags, metadata, etc.) used in the different Slovenian language data resources we are dealing with. Just to give an example, we map the morpho-syntactic descriptors used in Sloleks to the standardized LexInfo representation, shown in Example 1, taken from the Python code realizing the transformation (not considering for the time being the gender information):

```
slo_lexinfo_table["Ncfsn"] = ("lexinfo:singular","lexinfo:nominativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfsg"] = ("lexinfo:singular","lexinfo:genitiveCase","lexinfo:feminine")
slo_lexinfo_table["Ncfsd"] = ("lexinfo:singular","lexinfo:dativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfsa"] = ("lexinfo:singular","lexinfo:accusativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfsl"] = ("lexinfo:singular","lexinfo:locativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfsi"] = ("lexinfo:singular","lexinfo:instrumentalCase","lexinfo:feminine")
slo_lexinfo_table["Ncfdn"] = ("lexinfo:dual","lexinfo:nominativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfdg"] = ("lexinfo:dual","lexinfo:genitiveCase","lexinfo:feminine")
slo_lexinfo_table["Ncfdd"] = ("lexinfo:dual","lexinfo:dativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfda"] = ("lexinfo:dual","lexinfo:accusativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfdl"] = ("lexinfo:dual","lexinfo:locativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfdi"] = ("lexinfo:dual","lexinfo:instrumentalCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpn"] = ("lexinfo:plural","lexinfo:nominativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpg"] = ("lexinfo:plural","lexinfo:genitiveCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpd"] = ("lexinfo:plural","lexinfo:dativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpa"] = ("lexinfo:plural","lexinfo:accusativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpl"] = ("lexinfo:plural","lexinfo:locativeCase","lexinfo:feminine")
slo_lexinfo_table["Ncfpi"] = ("lexinfo:plural","lexinfo:instrumentalCase","lexinfo:feminine")
```

Example 1: Mapping the abbreviated morpho-syntactic code of Sloleks to the LexInfo vocabulary, which is equipped with unique reference points (URIs), as the name space "lexinfo" is standing for: "http://www.lexinfo.net/ontology/3.0/lexinfo"

4 The integrated Slovenian Language Data in the OntoLex-Lemon Representation

In this phase of our work we focused on nouns. From Sloleks, 50,873 nouns have been transformed onto instances of the ontolex:LexicalEntry class, while 854,813 morphological variants are now encoded as instances of the ontolex:Form class. The entries are declaratively linked to the forms via the properties ontolex:canonicalForm (linking the entry to its singular nominative form) or ontolex:otherForm (linking the entry to all other form variants). The forms are encoding the number information (while the gender and the pronunciation information are still to be added).

We have also mapped 18,735 sloWNet nominal entries onto the LexicalConcept class of OntoLex-Lemon. Some nominal entries included in sloWNet could not be considered, as we could not map their lemmas to Sloleks entries. This concerns mainly multiword items, which are not considered in Sloleks, but which might be present in SLD. This is a topic which stays next on our agenda. The sloWNet concepts, now encoded as part of a SKOS scheme, are linked to the Sloleks entries via the ontolex:isEvokedBy property (on the other way round by the ontolex:evokes property).

This way, both original resources, Sloleks and sloWNet are interlinked and merged in one and the same representation space. A benefit is for example that all form variants of Sloleks are now related to a sloWNet element.

In the following tables, we display the resulting representation for *alergija* at the lexical, morphological, and conceptual levels. In the first table (Table 5) the lexical entry representation is displayed. In Table 6 few examples of the form variants the entry is linking to are displayed (as the rdfs:label property is widely used in the Linked Data community, we keep it in parallel to the ontolex:writtenRep property). In Table 7 we can see the corresponding synset of sloWNet.

```
:Lex_646
  rdf:type ontolex:LexicalEntry ;
  rdfs:label "alergija"@slv ;
  ontolex:canonicalForm :Form_646_0 ;
  ontolex:evokes :eng-30-14533796-n ;
  ontolex:otherForm :Form_646_1 ;
  ontolex:otherForm :Form_646_10 ;
  ontolex:otherForm :Form_646_11 ;
  ontolex:otherForm :Form_646_12 ;
  ontolex:otherForm :Form_646_13 ;
  ontolex:otherForm :Form_646_14 ;
  ontolex:otherForm :Form_646_15 ;
  ontolex:otherForm :Form_646_16 ;
  ontolex:otherForm :Form_646_17 ;
  ontolex:otherForm :Form_646_2 ;
  ontolex:otherForm :Form_646_3 ;
  ontolex:otherForm :Form_646_4 ;
  ontolex:otherForm :Form_646_5 ;
  ontolex:otherForm :Form_646_6 ;
  ontolex:otherForm :Form_646_7 ;
  ontolex:otherForm :Form_646_8 ;
  ontolex:otherForm :Form_646_9 ;
.
```

Table 5: the OntoLex-Lemon representation of the entry *alergija*, with internal links to the forma variants and to the related sloWNet concept

```
:Form_646_10
  rdf:type ontolex:Form ;
  lexinfo:case lexinfo:locativeCase ;
  lexinfo:number lexinfo:dual ;
  rdfs:label "alergijah"@slv ;
  ontolex:writtenRep "alergijah"@slv ;
.
```

```
:Form_646_12
  rdf:type ontolex:Form ;
  lexinfo:case lexinfo:nominativeCase ;
  lexinfo:number lexinfo:plural ;
  rdfs:label "alergije"@slv ;
  ontolex:writtenRep "alergije"@slv ;
.
```

Table 6: Two example of form variant for the Sloleks entry *alergija*
Information on gender and pronunciation will be added soon.

```
:eng-30-14533796-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :SlowNet ;
  ontolex:isEvokedBy :Lex_33897 ;
  ontolex:isEvokedBy :Lex_646 ;
.
```

Table 7: The sloWNet synset corresponding to the entry *alergija*. This synset is also evoked by another lexical entry.

The Tables 5-7 show how Sloleks and sloWNet are now in the same (harmonized) representation space and how they enrich each other.

While working with sloWNet, we noticed that only (if at all) very few definitions (glosses) and examples in Slovenian language are associated with the synsets. This a reason why we also started to work with the data included in SLD, as there a richer combination of semantic and syntactic aspects is described, also with links to corpora data. Our goal is then to extract such examples and definitions and to associate those with the sloWNet synsets, but also with the morphological forms which are included in the examples.

A first step towards this goal was to extract all examples and to be able to encode them in the OntoLex-Lemon environment. We defined for this a class “Examples” and we store examples associated with an entry as instances of this class, as this is displayed in Table 8 below.

```
:Example_45
  rdf:type :Examples ;
  rdfs:comment "for the lemma alergija"@en ;
  rdfs:label " Zdravljenje <i>alergij</i> se zadnja leta izboljšuje z boljšimi sredstvi za diagnosticiranje, kot so krvni in kožni testi "@sl .

:Example_46
  rdf:type :Examples ;
  rdfs:comment "for the lemma alergija"@en ;
  rdfs:label " Znaki <i>alergije</i> pa so predvsem odvisni od mesta in organa, kjer se začne alergična reakcija. "@sl .

:Example_47
  rdf:type :Examples ;
  rdfs:comment "for the lemma alergija"@en ;
  rdfs:label "Najbolj običajni simptomi <i>alergije</i> so izpuščaji na koži in driska. "@sl .
```

Table 8: Example of the current implementation of the integration of examples taken from SLD.

Current work is being pursued in disambiguated some of the examples included for their appropriate linking to the concepts. As the reader can see, we still have for now tags around the form variants, as we want to extract those for relating the examples not only to the lemma, but to the concrete forms.

SLD is also containing a very rich list of collocations (the example sentences are in fact examples of such collocations, as encountered in large corpora). Current work consists in representing those collocations in OntoLex-Lemon. We started for this also a discussion within the W3C Community Group "Ontolex". An option would be to interlink the form variants that are in a collocation relation. Another option would be to consider collocations as kind of multiple word expressions. We are consulting for this also an additional Slovene lexical resource, MWELex, described in (Ljubesic et al. 2015), which also has a cross-lingual perspective. Representing collocations as MWE in OntoLex-Lemon can be done with the help of its decomp module.

```
:MWE_diagnoza_alergije
  rdf:type ontolex:MultiWordExpression ;
  rdfs:label "diagnoza alergije"@slv ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :alergije_comp ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :diagnoza_comp ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :Lex_6400 ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :Lex_646 ;
.
:MWE_ sprožilec _alergije
  rdf:type ontolex:MultiWordExpression ;
  rdfs:label " sprožilec alergije"@slv ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :alergije_comp ;
  <http://www.w3.org/ns/lemon/decomp#constituent> : sprožilec _comp ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :Lex_41850 ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :Lex_646 ;
.
:alergije_comp
  rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
  lexinfo:case lexinfo:genitiveCase ;
  lexinfo:number lexinfo:singular ;
  rdfs:label "alergije"@slv ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_646 ;
.
:diagnoza_comp
  rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
  rdfs:label "diagnoza"@slv ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_41850 ;
.
: sprožilec _comp
  rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
  rdfs:label " sprožilec "@slv ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_6400 ;
.
```

Table 9: The Lemon-OntoLex encoding of the collocations "diagnoza alergije" and "sprožilec alergije"

Table 9 shows how the components of the MWE expressions are linked to the lexical entries, and so indirectly to the corresponding sloWNet entries (as can be seen in Table 5). This opens an interesting field of investigations, as we could aim this way at generating new sloWNet entries based on the semantic composition of the components of the collocations.

5 Current and future Work

Current work is dedicated to the extension of the number and types of Slovenian data sets. Therefore, we started the analysis of terminological resources, as the relevance of OntoLex-Lemon for representing the lexical elements of a terminology has also been shown in (Cimiano et al. 2015). We consider here the Slovene terminology "Evroterm".

The downloadable version of Evroterm mentions 130.000 terms (for the date 10.10.2017). It is containing a mix of information, but not making a proper use of XML syntax. This aspect can be solved relatively easily, and we already mapped the data into a clean and machine interpretable XML structure. This resource is relevant to our project, as it also contains definitions in Slovene language that could be linked to the sloWNet data, which are lacking definitions in the Slovene language. Evroterm also contains term equivalents in various languages.

No part-of-speech information is assigned to the expressions realizing the terms. This is something we can easily add by linking the terms to Sloleks, at least for the terms that are not realized by multi-word-expressions but by sole words. Table 10 just below shows how the word *alergija* is encoded in the downloadable version of Evroterm.

```
<Entry Number>75845
<Subject>medicine ME
<Subj>medicina
<EN>allergy
<Definition>A condition of abnormal sensitivity in certain individuals to contact with substances such as proteins, pollens, bacteria, and certain foods. This contact may result in exaggerated physiologic responses such as hay fever, asthma, and in severe enough situations, anaphylactic shock.
<DefRef>KOREN
<SL>alergija
<TermRef>Besednjak Gemet - http://eionet-si.arso.gov.si/kpv/Gemet
<Definition>Nenormalna obèutljivost pri nekaterih posameznikih na stik z doloèenimi snovmi, npr. beljakovinami, cvetnim prahom, bakterijami in doloèeno vrsto hrane. Ta stik lahko privede do pretirane fiziološke reakcije, kot je seneni nahod, astma ter v težkih primerih tudi do anafilaktiènega šoka.
<DefRef>KOREN
<DA>allergi
<CS>alergie
<DE>Allergie
<ES>alergia
<FI>allergia
<FR>allergie
<IT>allergia
<NL>allergie
<PL>alergia
<Definition>swoista reakcja ustroju na pewne zwi¹zki chemiczne znajduj¹ce siê m.in. w powietrzu, w pokarmach, w bakteriach, bêd¹ca przyczyn¹ wielu chorób, np. astmy, pokrzywki i i
<PT>alergias
<SK>alergia
<SV>allergi
```

Table 10: The representation of the term *alergija* in Evroterm

The cross-lingual aspects present in Evroterm can be dealt with in OntoLex-Lemon with the help of its vartrans module. A cross-lingual resources, called MWELex, for multiple word expressions is also described in (Ljubesic et al. 2015), where MWEs are aligned across Serbian, Croatian, and Slovenian. The transformation of this resource onto OntoLex-Lemon and its linking to the other transformed and integrated Slovenian resources is also part of our current work.

6 Conclusions

We presented on-going work consisting in transforming a series of different rich Slovene lexical resources onto the OntoLex-Lemon framework. The current state of work shows the benefits of such an approach, as the lexical information from different resources is now available in one and the same representation format, supporting their interlinking and even merging.

The next steps will consist of representing the collocation information included in the cross-lingual MWELex resource and correctly linking it to elements of the current OntoLex-Lemon integrated data set. We will also continue working on the Evroterm data, as this resource is not only providing for terminological knowledge, but also with term translations.

7 References

Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics First Look*. 52 pages.

Cimiano, P., McCrae, J. & Buitelaar, P. (2016). Lexicon Model for Ontologies. *W3C Community Report*.

Cimiano, P., McCrae, J., Rodriguez-Doncel, V., Gornostaya, T., Gomez-Perez, A., Siemoneit, B., & Lagzdins, A. (2015). Linked Terminology: Applying Linked Data Principles to Terminological Resources. In *Proceedings of eLex 2015*.

Declerck, T., McCrae, J., Navigli, R., Zaytseva, K. & Wissik, T. (2018). ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data. In *Proceedings of the 2nd GLOBALEX Workshop.*

Declerck, T., Tiberius, C. & Wandl-Vogt, E. (2017). Encoding lexicographic Data in lemon: Lessons learned. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. Galway, Ireland, CEURS, 8/2017

Dobrovoljc, K., Krek, S. & Erjavec, T. (2017). The Sloleks Morphological Lexicon and its Future Development. In *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana University Press, Faculty of Arts.

Evroterm. Accessed at https://evroterm.vlada.si/evroterm. [31/05/2020]. The terminology data in text format can be downloaded at: http://podatki.vlada.si/evroterm.aspx [31/05/2020].

Fišer, D., Novak, J. & Erjavec, T. (2012). sloWNet 3.0: development, extension and cleaning. In *Proceedings of the 6th International Global Wordnet Conference*. The Global WordNet Association.

Fišer, D., Gantar, P. & Krek, S. (2012). Using explicitly and implicitly encoded semantic relations to map Slovene wordnet and Slovene lexical database. In *Proceedings of the 8th Conference on Language Resources and Evaluation*.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel., Pet,M. & Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Gantar, P. & Krek, S. (2011). Slovene lexical database. In *Proceedings of the sixth international Conference on Natural language processing, multilinguality.*

Krek, S., Kosem, I., McCrae, J.P., Navigli, R., Pedersen, B.S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts.*

Lexicon Model for Ontologies. Accessed at https://www.w3.org/2016/05/ontolex/ [31/05/2020].

LexInfo. Accessed at https://lexinfo.net/ontology/3.0/lexinfo [31/05/2020].

Ljubesic, N., Dobrovoljc, K., & Fiser, D. (2015). *MWELex - MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica (Slovenia), 39.*

Logar Berginc, N. & Kosem, I. (2011): Gigafida – the new corpus of modern Slovene: what is really in there? In *Proceedings of the Slavicorp conference*. Dubrovnik.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E. , Spohr, D. & Wunner, T. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

Morphological lexicon Sloleks 2.0. Accessed at: https://www.clarin.si/repository/xmlui/handle/11356/1230 [31/05/2020]

Racioppa, S. & Declerck, T. (2019). Enriching Open Multilingual Wordnets with Morphological Features. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari, Italy, CEUR, 10/2019

Semantic lexicon of Slovene sloWNet. Accessed at https://www.clarin.si/repository/xmlui/handle/11356/1026 [31/05/2020].

Slovene Lexical Database. Accessed at http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza [31/05/2020].

Tiberius, C. & Declerck, T. (2017). A lemon Model for the ANW Dictionary. In *Proceedings of the eLex 2017 conference, Pages 237-251.* Leiden, Netherlands, Lexical Computing CZ s.r.o., INT, Trojína and Lexical Computing, Brno, Czech Republic

Acknowledgements

Work by DFKI was supported by the European Union’s Horizon 2020 research and innovation program under grant agreement no. 825182 through the Prêt-à-LLOD project, and the COST Action CA18209, NexusLinguarum: European network for Web-centred linguistic data science. Work by the Austrian Centre for Digital Humanities and Cultural Heritage was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 through the ELEXIS project. We would also like to thank the anonymous reviewers for their helpful comments on the paper.