International Journal of Lexicography, 2019, Vol. 32, No. 2, 207–227
doi: 10.1093/ijl/ecy024
Advance Access Publication Date: 13 December 2018
Article

Article

CROATIAN NETWORK LEXICON WITHIN THE SYNTACTIC AND SEMANTIC FRAMEWORK AND LLOD CLOUD

Marko Orešković
National and University Library in Zagreb, Croatia (moreskovic@nsk.hr)

Sandra Lovrenčić
Faculty of Organisation and Informatics, Varaždin, Croatia (sandra.lovrencic@foi.hr)

Mario Essert
Faculty of Mechanical Engineering and Naval Architecture, Zagreb, Croatia (messert@fsb.hr)

Abstract

This paper presents a new type of network lexicon for the Croatian language based on a syntactic and semantic computational framework. It begins with an overview of the existing Croatian e-dictionaries and online repositories, as well as a brief outline of other relevant network ontological models. The network lexicon, which is based on an innovative approach to word tagging, is described in the remainder of the paper. Instead of presenting a linear (e.g. MULTEX-East) structure, this paper proposes a new hierarchical tree-like T-structure that is very similar to the structure of an ontology. In this approach, each word is processed on multiple levels: from its internal structure (morphs or syllables), via links to external network resources (encyclopaedias), to multiword expressions that can have distinctive roles, such as semantic domains, collocations and even figurative expressions. A network framework facilitates the fetching and filtering of the information related to the searched word in a paradigmatic sense because of the integration of the CroWN, the Croatian version of the English WordNet, and in a syntagmatic sense by building the database of the T-structure patterns from a selected corpus. Finally, the network framework enables the dynamic integration of the lexicon with the Linguistic Linked Open Data cloud; thus, each change in the lexicon will be automatically reflected in the cloud. It is therefore not necessary to perform any periodical synchronisation of the data, a task that is quite common when working with triples stored in a Virtuoso database. Special attention has been paid to the technical components and the data preparation process, which are described in detail to serve as a guide for transforming existing lexicographic data into Linked Open Data triples.

1. Introduction

From a computational perspective, tagging can be viewed as a process of assigning tags (spoken or written) to concepts in the human mind. A tag can be vocalised (using a sequence of agreed sounds), written (using a sequence of agreed letters), drawn or gestured (using agreed symbolic gestures or drawings). The diversity of tagging marks is determined and limited by an individual's embodiment, senses (hearing, vision, touch, etc.) and skills (reading, writing, etc.). The purpose of tagging is to transfer a message to another person: to communicate. To ensure that the recipient of the message is able to understand it, the individual must already know the content of the tag. Alternatively, based on his or her previous sensory and mental experiences, the individual must be able to recognise the additional content with which the tag is associated. The performance of many of these procedures was greatly facilitated with the advent of computers in the mid-twentieth century. Electronic methods gradually replaced mechanical ones, revolutionising lexicography and bringing it into a completely new era: e-lexicography. However, an essential part of the lexicographic process remains impervious to technological advances and media availability. Each lexicographic process includes the stages of: a) collecting and storing data, b) finding and interconnecting data to form information and c) using this information to structure, to model and to publish knowledge.

In the early days of e-lexicography, computers were used only for minor assistance in classical data processing, primarily for collecting and storing data onto faster and cheaper media with larger capacities. With the technological advancements in hardware and software, computational methods are increasingly being used for information searches and machine-aided learning. Machine learning is also used in corpus analysis (the collection of samples from unstructured text and synthesis) and ontology building. According to the philosophical tradition, ontology represents the science of reality: the types and structures of objects, their properties and relationships, and the events that emerge from these relationships (Smith and Welty 2001). Ontology has been a subject of (philosophical) study long before the computer era, and there are disagreements within the computational sciences about the use of the term. Computer ontology, defined by Gruber (1993) as an explicit specification of the conceptualisation of a particular domain in any area, provides the highest level of machine storage and knowledge processing. In formal and machine-readable ontologies, knowledge is defined by classes and their occurrences: objects with their data properties and relations. The creation of new knowledge from existing ontology seems to be the most important characteristic of its creation (Antoniou et al. 2012), and this is the focus of this paper.

A network thesaurus for the Croatian language was created (see Orešković et al. 2016b) based on theoretical research of the Explanatory Combinatorial Dictionary (ECD) (Mel'čuk 2006) and the Generative Lexicon (Pustejovsky 1991). This paper presents the lexicographic structure and the directions for the network implementation of this online thesaurus. It also briefly reflects on the possible applications (Orešković et al. 2017). In Section 2, an overview of existing Croatian e-dictionaries, online repositories and other network relevant ontological models is presented. Section 3 describes an innovative approach to word tagging within a network lexicon that ontologically (taxonomy of data and information links) connects the morphosyntactics and semantics of a word and associates the word with its components (morphs and syllables) and its place in related multiword expressions (MWEs) (colocations, idioms, etc.). In Section 4, the role of the network lexicon within the general Syntactic and Semantic Framework[1] (SSF) (cf. Orešković et al. 2016c), as well as its possible applications in syntactic and semantic analyses, is presented. Section 5 details the process for integrating the thesaurus into a linked data[2] cloud. Section 6 contains the conclusions and outlook.

2. Background

2.1. An overview of existing lexicographic resources for the Croatian language

Croatian lexicography has a remarkably long and rich tradition. For more than five centuries, it has been an important part of the European lexicography built on the tradition of Latin dictionaries (Štrkalj Despot and Möhrs 2015). However, as these authors further indicate: 'After the important change of lexicographic paradigm brought by the era of e-dictionaries … the discontinuation of this tradition became very apparent, primarily in the number and quality of e-dictionaries compared to other European, and even other Slavic languages'. A great deal of effort is currently being invested into closing this gap, and some positive changes are already apparent. However, some crucial changes, especially in the information technology, are yet to occur (Orešković et al. 2016a).

This section of the paper provides an overview of the current state of the art of the Croatian e-lexicography based on Štrkalj Despot and Möhrs (2015), who have proposed the classification of contemporary Croatian lexicographic achievements. They have classified the contemporary Croatian lexicography as follows:

1. Corpus-driven dictionaries: for example, the *Croatian Frequency Dictionary* (Moguš et al. 1999) and the *Dictionary of Marulić's Judita* (Moguš 2001).
2. Corpus-based dictionaries: for example, the *Dictionary of the Croatian Language for Schools* (Birtić et al. 2012).
3. Dictionaries that are available in closed digital formats (CDs or DVDs, code): the *First Croatian School Dictionary* (Čilaš Šimpraga et al. 2008) and the *Big Dictionary of the Croatian Language* (Jojić 2015)
4. Open-access network dictionaries: Croatian linguistic portal[3]
5. Open-access lexical databases (including terminological databases): for example, CroWN[4], MetaNet.HR[5], e-Glava[6], CroVallex[7], Struna[8] and HRANA[9]
6. Dictionary portals (including terminological portals): Portals for the Croatian Lexicographic Heritage[10] and The MREŽNIK project[11], and the Croatian Terminological Portal[12]

In this paper, some Croatian encyclopaedic endeavours, primarily the Croatian encyclopaedia, are given the same level of importance as lexicographic resources. The *Croatian Encyclopaedia* was conceived at the Lexicographic Institute Miroslav Krleža (hereafter referred to as LZMK) in the early 1990s as a traditional print encyclopaedia (Jecić et al. 2016). The online *Croatian Encyclopaedia* is based on the printed edition, which was published in 11 volumes from 1999 to 2009. The online edition contains ∼70,000 articles published by 1,100 authors. It collects information from various smaller network lexicons, such as the Croatian Family Lexicon, the Movie Lexicon, the Football Lexicon and the Biographical Lexicon.

2.2. Shortcomings of existing lexicographic resources

A majority of the above-mentioned resources share similar shortcomings. Primarily, they work on the same computationally outdated principle (Parker 2008). There is the form for the search query, which, if found, returns the description, content, definition or term explanation with the field or area in which the term appears.

Other concerns regarding these repositories include the following:

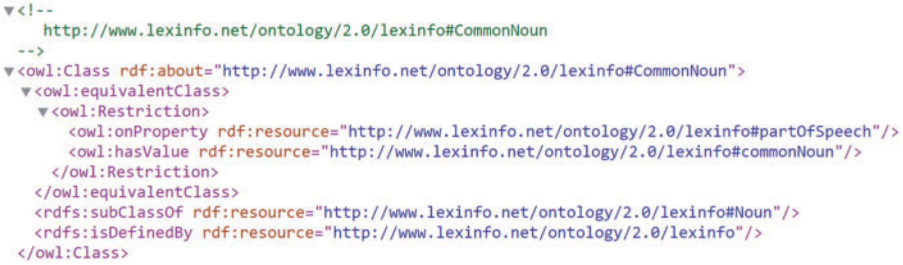
1. There is no connection between a word in the lexicon and the word in the real text for which the user needs help.
2. There is no connection between the words in the definitions and the words (entries) in the database. Definitions, especially short ones, may be incomprehensible as clues and may require a number of iterations until the true meaning of the concept is revealed.
3. Users cannot find a definition in the specialised dictionary if they do not know the clue word in advance. This works only if the users know the clue word and search for its definition. If they know the broad definition but need the correct form of the clue word, it does not work. This problem may be partly solved if users guess the starting letters of the word and wait for 'autocomplete' to list words starting with these letters; however, this rarely helps.
4. These repositories are hardly adaptable to rapid change. Centralised information collection without proper online access for entering and updating information, unlike Wikipedia's approach, makes these repositories outdated.
5. The information is not prepared for Linked Open Data (LOD) except in MetaNet.HR, which is partly prepared for the LOD environment but is still not part of the global Linguistic Linked Open Data (LLOD) cloud; thus, it cannot be linked to other similar repositories around the world.

From this overview, it can be easily deduced that there is a need for innovative approaches for building online solutions that will successfully overcome the above-mentioned deficiencies of the existing resources. Moreover, the new solutions should be able to connect the information in all of these resources automatically, allowing users to see all of the existing online information for each word they encounter in a text. This paper presents one such solution.

2.3. Linguistics ontologies

According to Prévot et al. (2010), ontologies and lexicons are representations of knowledge. However, Gruber (1993) distinguishes between formal and lexical ontologies, the latter conceptualising language resources based on linguistic criteria only and tending to have fewer formal specifications. Nevertheless, semantic technologies that currently incorporate ontologies and enable the publishing of information as linked data have found applications in linguistics and lexicography. Upper-level ontologies[13], such as CYC[14], the Suggested Upper Merged Ontology (SUMO)[15] or the Basic Formal Ontology (BFO)[16], could be of interest to the LLOD community. Various semantic dictionaries have been used in the development of linked data lexicons, such as XML, RDF, RDFS and OWL (see the W3C semantic web standards page for specifications and other details).[17] The benefits of the LOD to lexicography have already been recognised in recent works related to multilingual dictionaries (Bosque-Gil et al. 2016), etymological and dialectal dictionaries (Declerck et al. 2015) and recent international e-lexicographic projects (McCracken 2015). The LLOD projects relevant to this study are described below.

The LexInfo model[18], which builds on the previously developed LingInfo and LexOnto models and Lexical Markup Framework (LMF), was introduced in 2009 for associating



```
▼<!--
    http://www.lexinfo.net/ontology/2.0/lexinfo#CommonNoun
  -->
▼<owl:Class rdf:about="http://www.lexinfo.net/ontology/2.0/lexinfo#CommonNoun">
  ▼<owl:equivalentClass>
    ▼<owl:Restriction>
        <owl:onProperty rdf:resource="http://www.lexinfo.net/ontology/2.0/lexinfo#partOfSpeech"/>
        <owl:hasValue rdf:resource="http://www.lexinfo.net/ontology/2.0/lexinfo#commonNoun"/>
      </owl:Restriction>
    </owl:equivalentClass>
    <rdfs:subClassOf rdf:resource="http://www.lexinfo.net/ontology/2.0/lexinfo#Noun"/>
    <rdfs:isDefinedBy rdf:resource="http://www.lexinfo.net/ontology/2.0/lexinfo"/>
  </owl:Class>
```

Figure 1. Part of LexInfo ontology in OWL.

linguistic information with ontologies. According to Buitelaar et al. (2009), this facilitated linguistic descriptions that cannot be achieved with RDFS and OWL. LingInfo, developed in RDFS, enables the connection of the linguistic information for terms to classes and properties in the ontology through the definition of LingInfo instances (representations of terms) for each class or property (Buitelaar et al. 2006). LexOnto, developed in OWL, enables the mapping of sub-categorisation frames onto complex ontological structures (Cimiano et al. 2007). The LMF is an ISO standard for the development of natural language processing (NLP) lexicons with multilingual data (Francopoulo et al. 2009).

In addition to meeting the general requirements for supporting multilingualism, accessibility and interoperability, the LexInfo model was developed to fulfil five requirements (Buitelaar et al. 2009): 1) the separation and independence between the linguistic and ontological levels, 2) the expression of information about linguistic realisation, 3) the possibility of modelling the morphological and syntactic decomposition of complex terms, 4) the capture of the syntactic behaviour of lexical elements, and 5) the specification of the meaning of linguistic constructions with respect to domain ontology. According to Cimiano et al. (2011), LexInfo, along with the Linguistic Information Repository (LIR), provided a framework for the development of a new model: lemon, the Lexicon Model for Ontologies. LexInfo currently provides data categories for lemon. Version 3.0 is currently under development. A part of the LexInfo version 2.0 ontology that defines the class ‘common noun’ in OWL can be seen in Figure 1.

A direct link from the SSF to the LexInfo ontology is made by connecting each open word class from the LexInfo namespace to the appropriate class in the SSF namespace (see Figure 2). The details of the mapping process are described in Section 5. Along with the LexInfo, the SSF can be linked to any other semantic resource, such as BabelNet, DBpedia or WordNet. Only the word within the SSF lexicon needs to be properly tagged with an *owl:sameAs* tag containing the URI to an external resource. An example of such a linking is described briefly in Section 3.

3. Three types of lexicons

WordNet (Fellbaum 1998), as a semantic hierarchical structure of synsets, influenced the development of new data structures, T-structures, for the morpho-syntactic and semantic markups of words within a thesaurus framework. In semantic terms, a T-structure includes the vertical (paradigmatic) components (WordNet, linguistic portals, encyclopaedias, etc.)

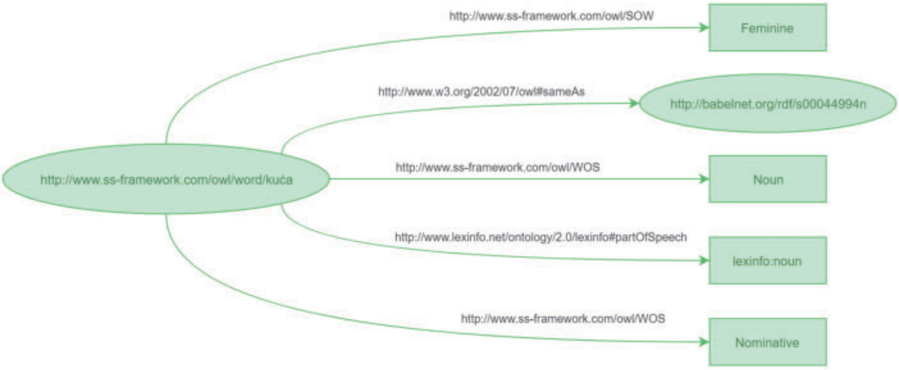

http://www.ss-framework.com/owl/SOW
Feminine
http://www.w3.org/2002/07/owl#sameAs
http://babelnet.org/rdf/s00044994n
http://www.ss-framework.com/owl/word/kuća
http://www.ss-framework.com/owl/WOS
Noun
http://www.lexinfo.net/ontology/2.0/lexinfo#partOfSpeech
lexinfo:noun
http://www.ss-framework.com/owl/WOS
Nominative

Figure 2. Link between the LexInfo part of speech and the SSF.

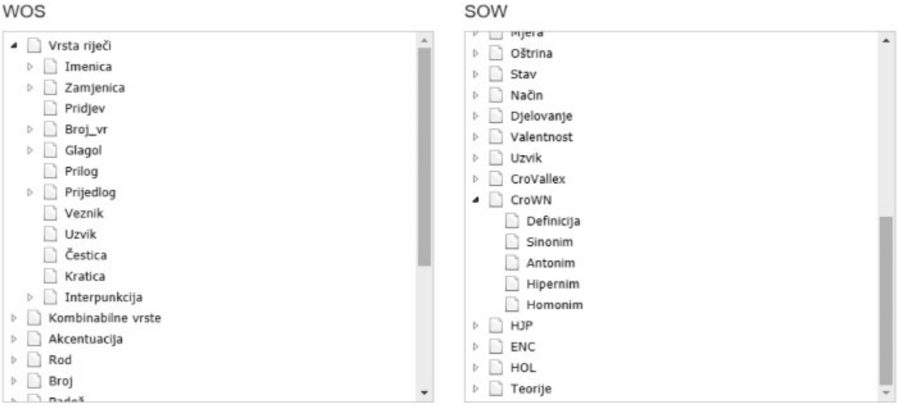
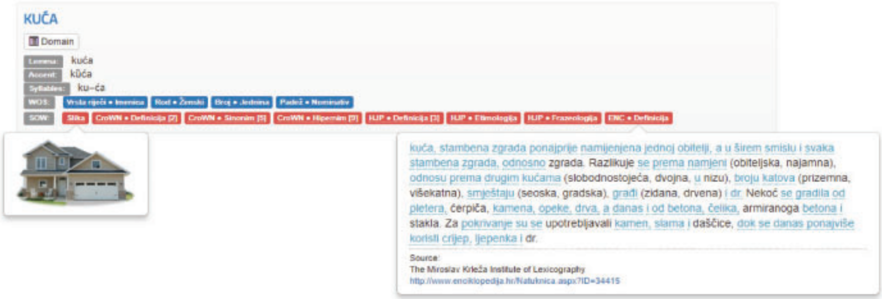


Figure 3. WOS and SOW tagging.

and the horizontal (syntagmatic) values for the subsequent building of collaborative databases. Thus, all definitions (glosses) in WordNet or any other encyclopaedia linked to the framework get their words linked, and this results in a significant expansion of the number of linked semantic nodes. Such structures are recursive and can be infinitely deep. To implement the T-structures in the SSF, this paper proposes two main categories: the word of speech (WOS), which contains information about the grammatical features of a word and is similar to the Part-of-speech (POS) but with a hierarchical structure, and the semantics of the word (SOW), which focuses on the word's semantic properties. Figure 3 shows a screenshot of the WOS and SOW tree for the Croatian language. The same applies to other languages. All users of the SSF can create their own WOS or SOW structures and assign them to words within the Lexicon. When the WOS or SOW structure of the deepest level is assigned to a word, all of the other categories on a higher level are also assigned.

A WOS or SOW tag can serve additional functions. It can have a specific value (strings, numbers, different identifiers, images or sounds). This is extremely useful in cases where



KUČA

Domain

Lemma: kuća
Accent: kȕća
Syllables: ku–ća
WOS: Vrsta riječi • Imenica | Rod • Ženski | Broj • Jednina | Padež • Nominativ
SOW: Slika | CroWN • Definicija [2] | CroWN • Sinonim [5] | CroWN • Hiperonim [9] | HJP • Definicija [3] | HJP • Etimologija | HJP • Frazeologija | ENC • Definicija

kuća, stambena zgrada ponajprije namijenjena jednoj obitelji, a u širem smislu i svaka stambena zgrada, odnosno zgrada. Razlikuje se prema namjeni (obiteljska, najamna), odnosu prema drugim kućama (slobodnostojeća, dvojna, u nizu), broju katova (prizemna, višekatna), smještaju (seoska, gradska), građi (zidana, drvena) i dr. Nekoć se gradila od pletera, čerpiča, kamena, opeke, drva, a danas i od betona, čelika, armiranoga betona i stakla. Za pokrivanje su se upotrebljavali kamen, slama i daščice, dok se danas ponajviše korisli crijep, ljepenka i dr.

Source:
The Miroslav Krleža Institute of Lexicography
http://www.enciklopedija.hr/Natuknica.aspx?ID=34415

Figure 4. WOS and SOW tags in the lexicon.

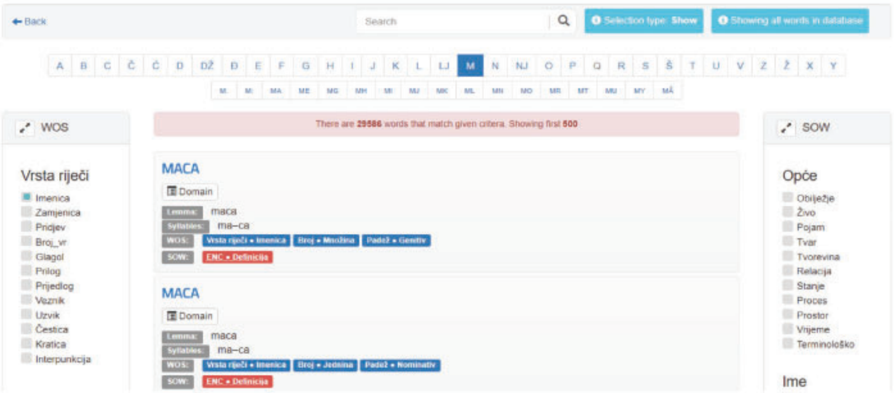
additional information, such as definitions from external resources like WordNet, the Miroslav Krleža Institute of Lexicography and the Croatian Language Portal, needs to be assigned to a word. Every definition within the SOW tags is automatically linked to the main Lexicon, which produces approximately ten times as many semantic relations. The lexicons are built from the corpora or external resources initially. Figure 4 shows one word from the SSF lexicon with a SOW tag containing image and textual values (definitions) for the word *kuća* (eng. 'house').

Sounds and images for words provide a solid foundation for any later lexicon usage in elementary school education. In addition to having WOS or SOW tags, every word in the lexicon within the SSF has other properties, such as lemma, syllable, morphs, and the appearance of the word in MWEs, assigned to it. There are three main types of lexicons in the SSF:

1. Lexicon of subatomic lexical entries (morphs, syllables or syllablemorphs). This lexicon contains ~2,000 morphs and ~8,000 syllables. These elements are used to build the entire word lexicon. It offers a great foundation for word formation research. Each morph or syllable is stored in a database as a unique element that enables the user to find only the words that have observed morphs or syllables on specific positions within the word itself.
2. Lexicon of words. Also known as the central lexicon, it currently contains ~800,000 words with the appropriate WOS or SOW tags. For every word, in addition to the WOS or SOW marks, there is information about its lemma, accentuation, syllables, morphs, etc. Words are used for building the MWE lexicon in similar ways that morphs and syllables are used to build a lexicon.
3. Lexicon of MWEs. This lexicon has ~120,000 MWEs that have been assigned WOS or SOW tags and are distributed across several meaningful groups, such as collocations, phrases and sets of synonyms. Each MWE, along with its corresponding WOS or SOW tags, is linked to words in a lexicon.

All three types of lexicons are organised in the same way. An example of crawling through the lexicons is shown in Figure 5. At the top of the screen is a search box for limiting the number of results.

Currently, the most common method for encoding word properties is the use of morphological tags. Every word in a dictionary has as many T-structure tags as possible. Not


← Back
Search
Selection type: Show
Showing all words in database
A B C Č Ć D DŽ Đ E F G H I J K L LJ M N NJ O P Q R S Š T U V Z Ž X Y
M. M: MA ME MG MH MI MJ MK ML MN MO MR MT MU MY MÁ
WOS
SOW
Domain

There are 29586 words that match given critera. Showing first 500

Vrsta riječi
- ■ Imenica
- □ Zamjenica
- □ Pridjev
- □ Broj_vr
- □ Glagol
- □ Prilog
- □ Prijedlog
- □ Veznik
- □ Uzvik
- □ Čestica
- □ Kratica
- □ Interpunkcija

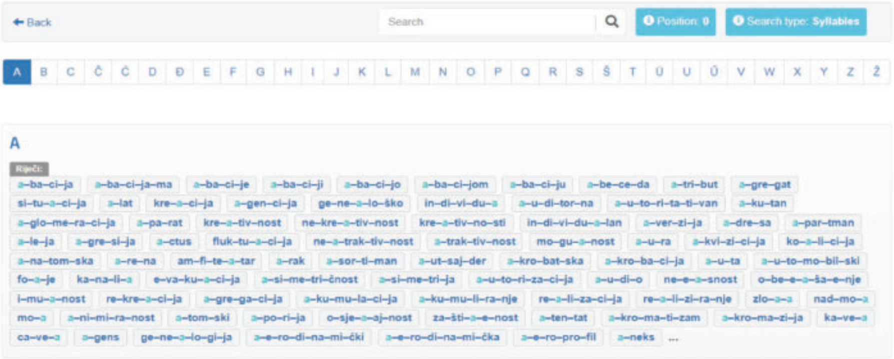
MACA
Lemma: maca
Syllables: ma–ca
WOS: Vrsta riječi • Imenica | Broj • Množina | Padež • Genitiv
SOW: ENC • Definicija

MACA
Lemma: maca
Syllables: ma–ca
WOS: Vrsta riječi • Imenica | Broj • Jednina | Padež • Nominativ
SOW: ENC • Definicija

Opće
- □ Obilježje
- □ Živo
- □ Pojam
- □ Tvar
- □ Tvorevina
- □ Relacija
- □ Stanje
- □ Proces
- □ Prostor
- □ Vrijeme
- □ Terminološko

Ime

Figure 5. Lexicon crawling.

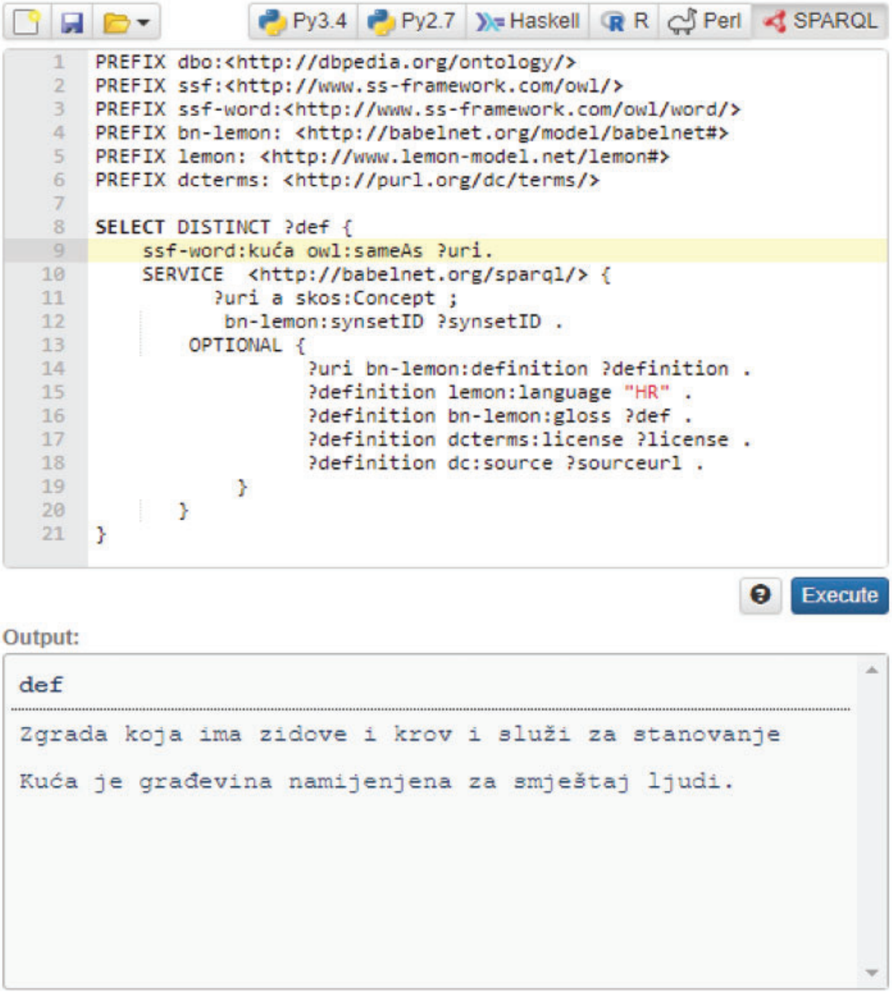

← Back
Search
Position: 0
Search type: Syllables
A B C Č Ć D Đ E F G H I J K L M N O P Q R S Š T Ü U Ö V W X Y Z Ž
A

Riječi:
a–ba–ci–ja | a–ba–ci–ja–ma | a–ba–ci–je | a–ba–ci–ji | a–ba–ci–jo | a–ba–ci–jom | a–ba–ci–ju | a–be–ce–da | a–tri–but | a–gre–gat | si–tu–a–ci–ja | a–lat | kre–a–ci–ja | a–gen–ci–ja | ge–ne–a–lo–ško | in–di–vi–du–a | a–u–di–tor–na | a–u–to–ri–ta–ti–van | a–ku–tan | a–glo–me–ra–ci–ja | a–pa–rat | kre–a–tiv–nost | ne–kre–a–tiv–nost | kre–a–tiv–no–sti | in–di–vi–du–a–lan | a–ver–zi–ja | a–dre–sa | a–par–tman | a–le–ja | a–gre–si–ja | a–ctus | fluk–tu–a–ci–ja | ne–a–trak–tiv–nost | a–trak–tiv–nost | mo–gu–a–nost | a–u–ra | a–kvi–zi–ci–ja | ko–a–li–ci–ja | a–na–tom–ska | a–re–na | am–fi–te–a–tar | a–rak | a–sor–ti–man | a–ut–saj–der | a–kro–bat–ska | a–kro–ba–ci–ja | a–u–ta | a–u–to–mo–bil–ski | fo–a–je | ka–na–li–a | e–va–ku–a–ci–ja | a–si–me–tri–čnost | a–si–me–tri–ja | a–u–to–ri–za–ci–ja | a–u–di–o | ne–e–a–snost | o–be–e–a–ša–e–nje | i–mu–a–nost | re–kre–a–ci–ja | a–gre–ga–ci–ja | a–ku–mu–la–ci–ja | a–ku–mu–li–ra–nje | re–a–li–za–ci–ja | re–a–li–zi–ra–nje | zlo–a–a | nad–mo–a | mo–a | a–ni–mi–ra–nost | a–tom–ski | a–po–ri–ja | o–sje–a–aj–nost | za–šti–a–e–nost | a–ten–tat | a–kro–ma–ti–zam | a–kro–ma–zi–ja | ka–ve–a | ca–ve–a | a–gens | ge–ne–a–lo–gi–ja | a–e–ro–di–na–mi–čki | a–e–ro–di–na–mi–čka | a–e–ro–pro–fil | a–neks ...

Figure 6. MSY lexicon.

only the words but also their components can be tagged (syllables or morphs). Because syllables and morphs are treated as word components, they can be considered sub-atomic components, and it is possible to construct a new type of a dictionary: the morphosyllabic dictionary. This creates new opportunities for the analysis of language at this lower 'sub-atomic' level (Figure 6). Such a lexicon can be filtered in many ways, such as a specific tag assigned to a word or the position of a morph or syllable within the word. If the same principle is applied in reverse, then words are components of a higher-order structure, the MWE. Like any other word, an MWE can be tagged with the proposed T-structures, resulting in the generation of an MWE dictionary. This dictionary can be further filtered by more specific criteria, such as the position of the word within the MWE, phrase, idiom or collocation, just as any other lexicon in the SSF.

Along with the harvested information, the URIs of external resources are stored in a SOW tag, *owl:sameAs*, which enables the SSF to be semantically connected to a global linguistic linked data (LLD) cloud. For example, the word hrv. *kuća* (eng. 'house') has a SOW tag *owl:sameAs* with the value http://babelnet.org/rdf/s00000356n, which corresponds to the same word in the BabelNet lexicon. This would now provide options for using the



```
PREFIX dbo:<http://dbpedia.org/ontology/>
PREFIX ssf:<http://www.ss-framework.com/owl/>
PREFIX ssf-word:<http://www.ss-framework.com/owl/word/>
PREFIX bn-lemon: <http://babelnet.org/model/babelnet#>
PREFIX lemon: <http://www.lemon-model.net/lemon#>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?def {
    ssf-word:kuća owl:sameAs ?uri.
    SERVICE  <http://babelnet.org/sparql/> {
          ?uri a skos:Concept ;
           bn-lemon:synsetID ?synsetID .
        OPTIONAL {
                    ?uri bn-lemon:definition ?definition .
                    ?definition lemon:language "HR" .
                    ?definition bn-lemon:gloss ?def .
                    ?definition dcterms:license ?license .
                    ?definition dc:source ?sourceurl .
             }
        }
}
```

```
def
Zgrada koja ima zidove i krov i služi za stanovanje
Kuća je građevina namijenjena za smještaj ljudi.
```

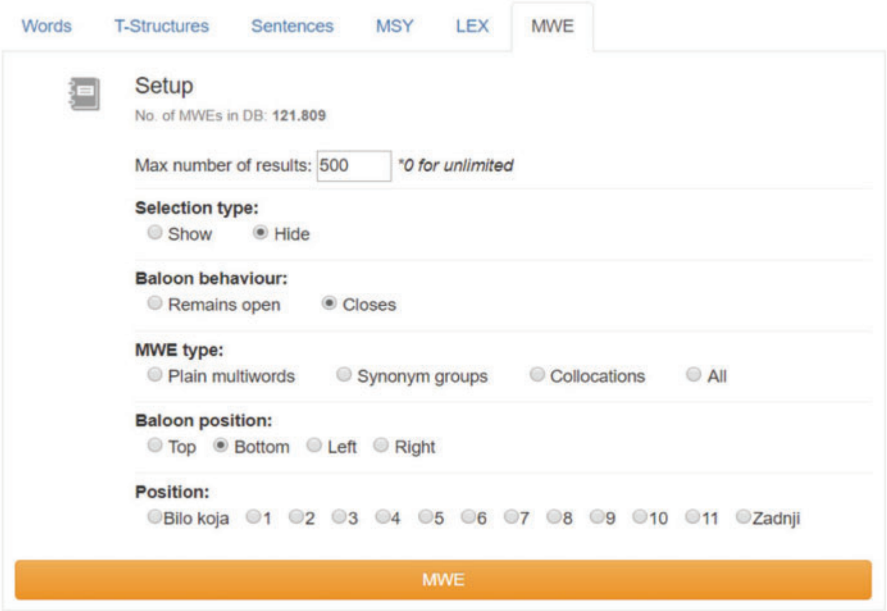
Figure 7. SPARQL query to retrieve definition from BabelNet.

information from BabelNet, thereby enriching the SSF lexicon. Figure 7 shows an example of a SPARQL query that uses the SSF lexicon. For the word hrv. *kuća*, it gets the definition from the BabelNet database.

Information can be retrieved from any other semantic resource in the same way. The only prerequisite is that the appropriate *owl:sameAs* SOW tag in the SSF be assigned.

4. Dictionary in the service of the SSF

A well-organised lexicon in which words are assigned as many tags as possible serves as a good foundation for the creation of a repository of sentence patterns. Because the words are assigned as many tags as possible, the number of different possible patterns is increased. Such a repository is built over the selected corpora by a pattern builder algorithm. In the



Setup

No. of MWEs in DB: 121.809

Max number of results: 500 *0 for unlimited

Selection type:
○ Show ● Hide

Baloon behaviour:
○ Remains open ● Closes

MWE type:
○ Plain multiwords ○ Synonym groups ○ Collocations ○ All

Baloon position:
○ Top ● Bottom ○ Left ○ Right

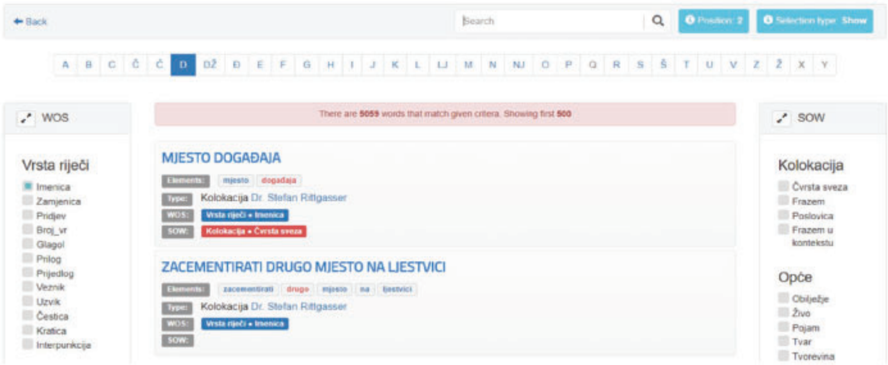
Position:
○ Bilo koja ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ 11 ○ Zadnji

Figure 8. MWE search settings.

SSF, the tags are assigned by using the above-mentioned T-structures, which are divided into two main categories. In this work, the SOW tags are used for semantic tagging, and the WOS words are used for grammatical tagging. The algorithm iterates over the corpora and creates a unique pattern for each sentence. Every word in a sentence is analysed in relation to selected WOS or SOW tags, and those that apply are included in the pattern. Once all the sentences have been processed, a set of unique patterns based only on the relevant word features will have been created. The SSF can later use these patterns to extract sentences that are similar or the same at the syntactic or semantic level.

The meanings of phrasemes provided in specialised dictionaries should be much more detailed than the meanings of single words in general dictionaries even in cases where the specific MWEs containing the word are provided (see Figure 5). Two Croatian dictionaries treat words in these ways: the *Croatian-English Dictionary of Phrasal Words and Idioms* (Vrgocč and Fink-Arsovski 2008), and the *Croatian Dictionary of Synonyms* (šarić and Wittschen 2010). The first contains 2,490 Croatian phrasemes and 6,442 English equivalents, and the latter contains approximately 10,000 sequences of synonyms. In the SSF, the dictionary of synonyms is used in the same way as the CroWN or LZMK information inside the SOW tags. The above-mentioned dictionary of idioms is not yet being used in the SSF, but preparations are being made for its inclusion. Currently, the MWE tab in the SSF uses MWEs that are automatically obtained from network encyclopaedias and checked using publicly available repositories, such as 'Sveze riječi – Wortverbindungen' by Stefan Rittgasser[19]).

The settings form (see Figure 8) allows the user to define the lexicon display. The MWE type (idioms, colocations, proverbs, etc.) and alphabetic order for a given word, with the WOS or SOW tags, according to its position in the MWE are very important parameters.



There are 5059 words that match given critera. Showing first 500

MJESTO DOGAĐAJA
Elements: mjesto dogadaja
Type: Kolokacija Dr. Stefan Ritlgasser
WOS: Vrsta riječi • Imenica
SOW: Kolokacija • Čvrsta sveza

ZACEMENTIRATI DRUGO MJESTO NA LJESTVICI
Elements: zacementirati drugo mjesto na ljestvici
Type: Kolokacija Dr. Stefan Ritlgasser
WOS: Vrsta riječi • Imenica
SOW:

Figure 9. Collocations with a noun ('događaja') and a number ('drugo') as the second element of the MWE.

For example, it is possible to search for a collocation that has a noun or a number as a second element (see Figure 9).

Such filtering is necessary because of the large number of MWEs. There are currently 121,775 MWEs in the SSF. To tackle this problem, it is possible to define the maximum number of results in the settings tab before performing a query. The default value is set to 500 to define more rigorous search criteria, such as a search for nouns, case or gender only. This problem is connected to the issues of linguistic classifications in lexicography (Atkins et al. 2008) and multiword combinations. Given that the SSF offers users a multifunctional dictionary, creating classes of multiword combinations, as was done in four Danish dictionaries (Bergenholtz and Gouws 2014) in which twenty-two classes were created, is necessary.

This will not be a difficult task because WOS tags can be easily expanded both vertically and horizontally. For example, a branch VERB can be subdivided into 'idiomatic' and 'non-idiomatic' sub-branches, which can be further subdivided into 'particle' and 'reflexive' sub-branches). This approach is more user-friendly and offers better visibility. Moreover, it can be successfully applied to other multiword combinations, such as similes, twin comparatives, formulas, or winged words. These are already covered in the SSF by patterns and domains that are used for detecting tropes (Orešković et al. 2017).

For many years, the modelling of lexicographic information has been moving toward the deep conversion of the lexicographic input structure from the list, a classic structure on which computer databases rely, to a graph of information dominated by links. In brief, lexicography is moving from the linear availability of stored content to linked information. WordNet (Fellbaum 1998) has already built its structure as a graph-based lexical-semantic database in which nodes represent synsets, sets of cognitive synonyms. A similar procedure, only not in lexical semantics but in the lexical system, was done in the French Lexical Network project (Polguère 2014), in which entries are viewed as part of a language system of related lexical elements, including paradigmatic and syntagmatic relationships drawn from the set of lexical functions of the meaning-text theory (Mel'čuk 1996). The result is a multi-dimensional graph with a wide range of relations linking the lexical elements in their nodes.

The LLOD cloud and the proposed models, such as LexInfo, GOLD, SKOS, and lemon, for converting resources into them do not represent a new way of developing lexicons. However, they may facilitate and accelerate the processing of the internal structure significantly by prioritising the interoperability of multiple sources. The adaptation of the SSF for inclusion in the global linked data world should therefore be seen in this context.

5. Integration of network thesaurus into the LOD cloud

The Resource Description Framework (RDF) is the formal backbone of interlinked resources whose elements are expressed as triples or statements of the subject-predicate-object form, where the subject and object are nodes of the resource elements and the predicate is the edge connecting the nodes. The result of this linking is a graph whose nodes are almost limitless. For example, they can be lexical units, morpho-syntactic tags, phrases or sentences. The main advantage of this approach is the semantic and syntactic interoperability provided by the RDF and the linguistic vocabularies (LexInfo or GOLD) that enables the integration, exchange and enrichment of lexicographic data among different resources and the reusability of the whole resource (Wandl-Vogt et al. 2015).

In the present work, the approach to this problem is to integrate all of the SSF data into a global linguistic cloud but only as an addition to an existing functionality. Thus, a Croatian lexicon could be retrieved in four different ways: a) through a visual interaction with a user, b) through programming language consoles (e.g. Python, R, and Haskell), c) through API functions for programmatically independent applications, and d) through SPARQL queries of the data stored in the Croatian Linguistic Linked Open Data (CroLLOD) cloud. Although this integrated approach is very demanding in terms of the complexity of the program realisations, it is the only one that enables the building of a network framework in which there is complete synchronism between diverse types of lexical data and the processing of these data through different applications. Thus, changes in the SSF relational database must simultaneously occur at the LOD endpoints, and subsequent (batch) updates of the data are not necessary. Specifically, instead of performing frequent updates, as with the Virtuoso data store that has triples that are modified in a relational database, it was necessary to create a layer (wrapper) over the relational data that transforms the data in the RDF triples directly, in real time, so that they are then crawlable and accessible via a SPARQL endpoint.

To integrate the Croatian language into a global cloud, semantic and syntactic operability must be provided. For the LLOD cloud, the ontology URIs: 1) should be resolvable, 2) should be in an RDF format (RDFa, RDF/XML, Turtle or N-Triples), 3) should contain at least 1,000 triples, 4) should be linked to other resources in the LLOD, 5) must be crawlable (via an RDF dump or a SPARQL endpoint), 6) must contain linguistic data, and 7) must be registered in the DataHub. There are two main ways in which any data from a relational database can be displayed in an RDF format. The first, a static approach, is based on periodic synchronisation of the data from the relational database with the RDF data management system (e.g. the Virtuoso Triple Store) as shown in Figure 10. The second, a dynamic approach, is based on mapping the data from the relational database to RDF triples using the D2RQ platform (Figure 11).

The static approach implies the development of the skeleton ontology for storing lexical data. Each ontology is made according to its purpose, which determines the degree of

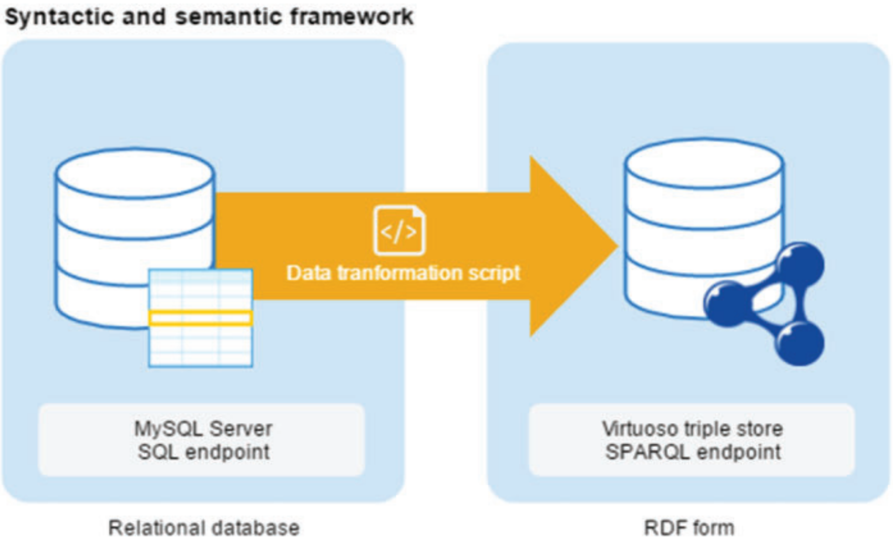


Figure 10. Static data synchronisation.

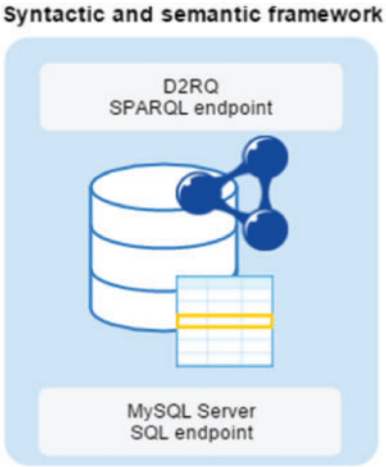


Figure 11. Dynamic data transformation.

complexity. One of the main ideas in building a linked data platform is that the data should be extendable at any time, not only in the authors' own ontology but also globally. The free ontology editor Protégé was used for base ontology model development. The class *word* holds information about dictionary words, the class *SOW* holds information about the semantic markups of the words (e.g. animate), and the class *WOS* holds information about the grammatical features of the words (e.g. nouns, verbs or adverbs). Although individuals can be defined in Protégé, only the ontology structure needed to be defined because individuals will later be inserted into the Virtuoso triple store directly by a special computer program. This will provide an ontology that is always synchronised with the dictionary that is part of the SFF.

After the ontology model is built, the second step toward achieving a fully functional linguistic ontology is the transfer of data from the relational database to the Virtuoso triple store. To accomplish this, it is necessary to develop the middleware application that transforms the data. This application uses SQL queries on a relational database and transforms returned datasets into triples that are then inserted into the Virtuoso triple store via a special computer program. The first step for a middleware application is to connect it to the MySQL database and iterate the whole dictionary. The computer program retrieves the WOS or SOW marks for each row in the *words* table. For each retrieved word, the Virtuoso server is contacted, and the RDF data for the selected word is updated. The biggest disadvantage of such an approach is that the data available through the SPARQL endpoints are not current copies of the relational database state. They denote the state that was valid at synchronisation.

Unlike the static approach, the dynamic approach can be conceived as a wrapper around the relational data. There are many means through which this form of transformation can be achieved, but one of the most common is the D2RQ platform. According to the official website, the 'D2RQ Platform is a system for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store'. Because all the data are created within the SSF, this read-only endpoint is suitable. The basic setup consists of the appropriate mapping file that tells the D2RQ server how to map relational database tables to the RDF. Figure 12 shows a part of the relational model with regard to the words and their WOS or SOW markups. Because of the many-to-many relationships, the data must be transformed into an RDF format. Although the whole database model of the SSF consists of more than seventy interrelated tables, only five are relevant for this paper. The table *words* contains the data for all the words in the dictionary, and the tables *wos* and *sow* are repositories for the WOS or SOW markups. They contain not only the global WOS or SOW marks but also the user-created markups, which can also be exported to a newly generated ontology. Because *word* can have one or more WOS or SOW marks and a WOS or SOW mark can be assigned to one or many *words*, the weak entities *word_has_wos* and *word_has_sow* have been introduced. They contain information about the WOS or SOW assigned to each word, as well as auxiliary data, such as the ordering of these markups. Within the SSF, these types of markups are known as T-structures because they can be easily represented as trees.

Instead of having the classic (e.g. MULTEXT-East) POS tagging of words for grammatical and some semantic categories (e.g. animate), these hierarchical T-structures can include various data types in their branches (strings, integers, links, word lists, ordered word lists, etc.), facilitating better descriptions of words and their various occurrence possibilities in a text. To transform these types of data into an RDF, the D2RQ server uses a 'mapping file' that describes how relational data will be presented in an RDF format and simultaneously enriches the lexical data by linking the WOS marks directly to a LexInfo ontology. The data interconnectivity is of immense importance for several applications, such as machine translation.

The D2RQ server has many advantages because the data from relational database do not need to be replicated to a triple store; however, it also has one important disadvantage. In the current version, there is still no support for executing federated SPARQL queries,

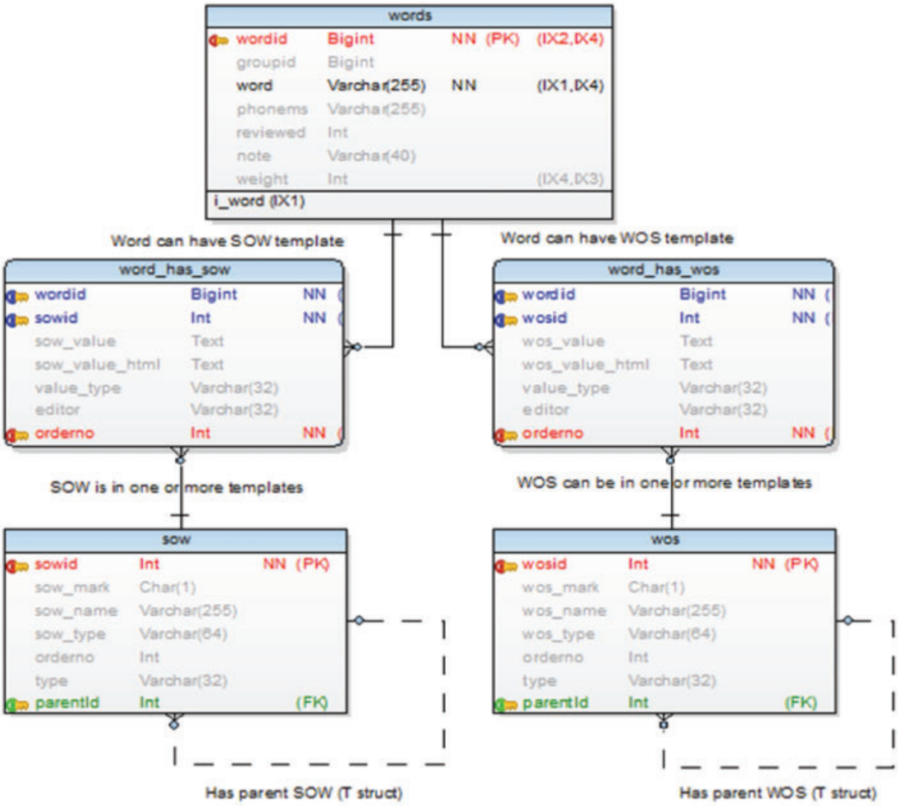


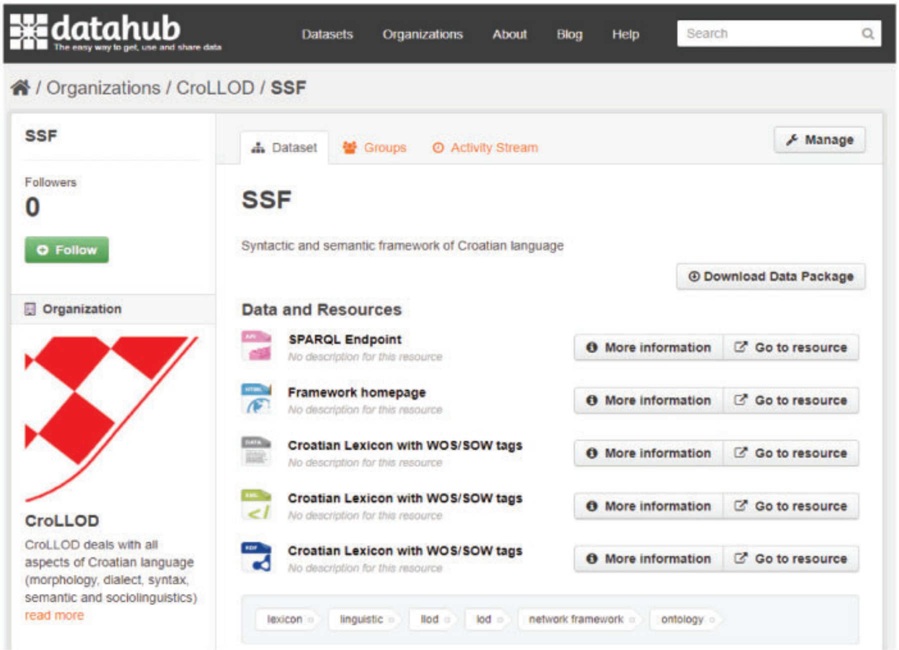
Figure 12. ER model of word-to-WOS and SOW relation.

thus limiting the D2RQ's utility to a local database. In contrast, the commercial edition of the Virtuoso Triple store allows both mapping to a relational database and the use of federated queries. The last request for LLOD inclusion is to register the datasets in the DataHub. The new organisation CroLLOD was created for the Croatian language.

As is shown in Figure 13, CroLLOD currently offers only a part of the SSF dictionary as either a downloadable ontology (in RDF, XML or N3 format) or over the SPARQL endpoint; thus, it is possible for a user to perform queries and retrieve only the relevant triples. In April 2018, the SSF Lexicon became part of the LOD cloud[20] with 70,366 triples, of which 67,717 are connected to LexInfo, 35,687 to the Princeton WordNet and 20,456 to BabelNet.

6. Evaluation

A majority of the words in the lexicon are validated manually. The remainder are created by the morphological generator by J. Markučič[21]. The morphological generator provides all of the grammatical tokens for an open word class in the Croatian language that respects all of the phonetical and morphological rules. The manual validation of each lemma in the lexicon is still expected to be done by Croatian language teachers or professors. The



/ Organizations / CroLLOD / SSF

SSF

Organization

CroLLOD

CroLLOD deals with all aspects of Croatian language (morphology, dialect, syntax, semantic and sociolinguistics)

SSF

Syntactic and semantic framework of Croatian language

Data and Resources

SPARQL Endpoint
No description for this resource

Framework homepage
No description for this resource

Croatian Lexicon with WOS/SOW tags
No description for this resource

Croatian Lexicon with WOS/SOW tags
No description for this resource

Croatian Lexicon with WOS/SOW tags
No description for this resource

lexicon linguistic llod lod network framework ontology

Figure 13. CroLLOD inside the DataHub.

validation project is led by the Department of Mathematics[22] in cooperation with the Faculty of Humanities and Social Sciences[23] at the University of Osijek. Another kind of word verification is done using the Natural Language Functions (NLF) functions *CountWOS()* and *CountSOW()*. These functions accept the WOS or SOW tag identification and can therefore return the number of words within the lexicon that have been assigned a specific tag. For example, through this process, it is known that the SSF lexicon contains 178,968 nouns, 117,421 verbs and 2,310 pronouns (in all grammatical forms). The third validation of the LLOD ontology that is included in the global LLOD cloud is done using the W3C RDF validator. After all these validations were completed, the SSF ontology was published in the DataHub[24]. When the new document is loaded into the SSF, there are two options if the parsed word is not in the SSF lexicon. The first is to insert the word into the lexicon and mark it as an unidentified linguistic object (ULO) so that a linguist, with the assistance of an administrator, can perform a manual or semi-automated check of the word. This would allow for the use of the morphological generator to create all of the tokens for the parsed word. The ULOs can be later classified and tagged to ensure that the lexicon is continuously improved. Thus, the SSF has complete control over the content in its lexicons.

7. Discussion and conclusions

This paper described the procedure for publishing a part of the Croatian lexicon (7% of ~800,000 words, ~120,000 MWE)[25] online. The three levels of procedures were explained:

1. The preparation of the lexicographic data by tagging with new T-structures, which is similar to the process for the ontological model, provides all of the grammatical and semantic properties of a word.
2. The preparation of the network lexicon as a part of the SSF enables word browsing and filtering using T-structures, smaller parts of the word (e.g. morphs and syllables) or MWEs that are related to the specific word. This is an excellent starting point for revolutionising further syntactic and semantic analyses of selected corpora.
3. The preparation of the data stored in a relational database for transformation to RDF triples is necessary for connecting the data to other nodes of the global linguistic cloud. However, a wrapper was used to avoid unnecessary data updates and synchronisation of the two sources.

In terms of linguistic technologies, the LLOD represents a powerful and relatively comfortable environment for the full integration of smaller and under-resourced languages into a global linguistic cloud. Lexicons conceived in this way offer new possibilities in NLP because they connect the syntactic and semantic levels with the morphological structures of words and their components.

There are three main possibilities for the future development of the SSF as part of a global network. The first is to expand the number of lexical entries that are available in the RDF format. Such an expansion will ensure that the complete Croatian dictionary (morphosyllables and MWEs) is available in the linguistic cloud. The second aspect is focused on creating links to other lexical ontologies besides LexInfo. This would certainly improve the visibility of Croatian, an under-resourced language, in a global linguistic cloud. Finally, the third aspect of the future work is to enrich the data provided in the linked data cloud with new datasets. The focus will be a database of sentence patterns similar to e-VALBU[26] and Erlangen Valency Patternbank[27]. These sentence patterns are a valuable resource for rule-based machine translation, the detection of plagiarism, the comparisons of similar languages (e.g. Croatian and Serbian), and a broad spectrum of syntactic and semantic analyses of textual documents.

Funding

This work was supported in part by the Croatian Science Foundation under the project IP-06-2016.

Acknowledgements

The authors thank the anonymous reviewers for their valuable comments and suggestions, which were helpful in improving the article.

Notes

[1] http://www.ss-framework.com

[2] The term *linked data* refers to a 'set of best practices for publishing and interlinking structured data on the Web' (Heath and Bizer 2011).

[3] http://hjp.novi-liber.hr/

[4] http://meta-share.ffzg.hr/repository/browse/croatian-wordnet

[5] http://ihjj.hr/metafore

6. http://valencije.ihjj.hr/
7. http://theta.ffzg.hr/crovallex
8. http://struna.ihjj.hr
9. http://hrana.ffzg.hr/
10. http://crodip.ffzg.hr/
11. http://ihjj.hr/projekt/hrvatski-mrezni-rjecnik-mreznik/70/
12. http://nazivlje.hr/
13. Upper-level ontologies contain general categories that are applicable across multiple domains.
14. http://www.cyc.com/
15. http://www.adampease.org/OP/
16. http://ifomis.uni-saarland.de/bfo/
17. https://www.w3.org/standards/semanticweb/ontology
18. http://lexinfo.net/index.html
19. http://www.lingua-hr.de
20. https://lod-cloud.net/dataset/ssf
21. https://messert.pythonanywhere.com/CROMorph
22. https://www.mathos.unios.hr/
23. http://www.ffos.unios.hr/
24. https://old.datahub.io
25. Complete lexicon currently contains over 800.000 of words whereas 7% of them are manually validated and as such suitable to be published in the global Linguistic Linked Open Data (LLOD) cloud.
26. http://hypermedia.ids-mannheim.de/evalbu/
27. http://www.patternbank.uni-erlangen.de

References

A. Dictionaries

Birtić, M., G. Blagus Bartolec, L. Hudeček, L. Jojić, B. Kovačević, K. Lewis, I. Matas Ivanković, M. Mihaljević, I. Miloš, E. Ramadanović and D. Vidović. 2012. *Školski Rječnik Hrvatskoga Jezika*. Zagreb: Školska knjiga: Institut za hrvatski jezik i jezikoslovlje.

Čilaš Šimpraga, A., L. Jojić and K. Lewis. 2008. *Prvi Školski Rječnik Hrvatskoga Jezika*. Zagreb: Školska knjiga: Institut za hrvatski jezik i jezikoslovlje.

Jojić, L. 2015. *Veliki Rječnik Hrvatskoga Standardnog Jezika*. Zagreb: Školska knjiga.

Moguš, M. 2001. *Rječnik Marulićeve Judite. (Rječnici hrvatskoga jezika 1.)*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Moguš, M., M. Bratanić and M. Tadić. 1999. *Hrvatski Čestotni Rječnik*. Zagreb: Školska knjiga.

Parker, P. M. (ed.). 2008. *Webster's Croatian-English Thesaurus Dictionary*. Las Vegas: ICON Group International.

Šarić, L. and W. Wittschen. 2010. *Rječnik Sinonima Hrvatskoga Jezika*. Zagreb: Jesenski i Turk.

Vrgoč, D. and Z. Fink-Arsovski. 2008. *Hrvatsko-Engleski Frazeološki Rječnik: Kazalo Engleskih i Hrvatskih Frazema = Croatian-English Dictionary of Idioms: Index of English and Croatian Idioms*. Zagreb: Naklada Ljevak.

B. Other literature

Antoniou, G., P. Groth, F. Van Harmelen and R. Hoekstra. 2012. *A Semantic Web Primer* (Third edition.). Cambridge: The MIT Press. Accessed on 14 April 2018. https://mitpress.ublish.com/ereader/147/?preview#page/Cover.

Atkins, B. T. S., T. S. Beryl and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press. Accessed on 2 October 2017. https://global.oup.com/academic/product/the-oxford-guide-to-practical-lexicography-9780199277711?cc=hr&lang=en&.

Bergenholtz, H. and R. Gouws. 2014. 'A Lexicographical Perspective on the Classification of Multiword Combinations.' *International Journal of Lexicography* 27.1: 1–24. Accessed on 2 October 2017. https://academic.oup.com/ijl/article-lookup/doi/10.1093/ijl/ect031.

Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda and G. Aguado-de-Cea. 2016. 'Modelling Multilingual Lexicographic Resources for the Web of Data: The K Dictionaries Case' In Kernerman, I., I. Kosem, S. Krek and L. Trap-Jensen (eds), *Proceedings of the GLOBALEX'16: Lexicographic Resources for Human Language Technology Workshop at the International Conference on Language Resources and Evaluation, LREC 2016, Portoroz, Slovenia, May 23-28, 2016*. Paris: European Language Resources Association, 65-72. Accessed on 4 October 2017. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf.

Buitelaar, P., P. Cimiano, P. Haase and M. Sintek. 2009. 'Towards Linguistically Grounded Ontologies' In Aroyo, L., P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou and E. Simperl (eds), *The Semantic Web: Research and Applications - 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 111–125. Accessed on 9 October 2017. https://pdfs.semanticscholar.org/1115/81d1335d836660c938ddfc854409d65dd27f.pdf.

Buitelaar, P., T. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos, V. Micelli, R. Porzel and P. Cimiano. 2006. 'Linginfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies', *Proceedings of the OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies Workshop at the International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*. Paris: European Language Resources Association, 28–34. Accessed on 4 October 2017. http://www.lrec-conf.org/proceedings/lrec2006/workshops/W14/Ontolex06.pdf

Cimiano, P., P. Buitelaar, J. McCrae and M. Sintek. 2011. 'LexInfo: A Declarative Model for the Lexicon-Ontology Interface.' *Web Semantics: Science, Services and Agents on the World Wide Web* 9.1: 29–51. Accessed on 5 October, 2017. http://www.sciencedirect.com/science/article/pii/S1570826810000892.

Cimiano, P., P. Haase, M. Herold, M. Mantel and P. Buitelaar. 2007. 'LexOnto: A Model for Ontology Lexicons for Ontology-Based NLP', *Proceedings of the OntoLex07- From Text to Knowledge: The Lexicon/Ontology Interface Workshop at the International Semantic Web Conference, ISWC'07, Busan, Korea, November 11–15, 2007*. Accessed on 4 October 2017. https://pdfs.semanticscholar.org/6486/d6d4ed9496b4f9ac4578d907d6314d6aaedc.pdf

Declerck, T., E. Wandl-Vogt and K. Mörth. 2015. 'Towards a Pan European Lexicography by Means of Linked (Open) Data' In Kosem, I., M. Jakubíček, J. Kallas and S. Krek (eds), *Electronic lexicography in the 21st century: linking lexical data in the digital age, Proceedings of the eLex 2015 conference, Herstmonceux Castle, United Kingdom, August 11-13, 2015*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, 342-

255. Accessed on 4 October, 2017. https://elex.link/elex2015/proceedings/eLex_2015_22_Declerck+etal.pdf.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Francopoulo, G., N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet and C. Soria. 2009. 'Multilingual Resources for NLP in the Lexical Markup Framework (LMF).' *Language Resources and Evaluation* 43.1: 57–70. Accessed on 9 October 2017. https://doi.org/10.1007/s10579-008-9077-5.

Gruber, T. 1993. 'A Translation Approach to Portable Ontology Specifications.' *Knowledge Acquisition* 5.2: 199–220.

Heath, T. and C. Bizer. 2011. 'Linked Data: Evolving the Web into a Global Data Space.' *Synthesis Lectures on the Semantic Web: Theory and Technology* 1.1: 1–136. Accessed on 14 April 2018. http://info.slis.indiana.edu/∼dingying/Teaching/S604/LODBook.pdf.

Jecić, Z., D. Boras and D. Domijan. 2016. 'Prilog Definiranju Pojma Virtualna Enciklopedija.' *Studia lexicographica* 1.2: 115–126. Accessed on 4 September 2017. http://studialexicographica.lzmk.hr/index.php/sl/article/view/32.

McCracken, J. 2015. 'The Exploitation of Dictionary Data and Metadata' In Durkin, P. (ed)., *The Oxford Handbook of Lexicography*. Oxford University Press. Accessed on 4 September 2017. http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199691630.001.0001/oxfordhb-9780199691630-e-36.

Mel'čuk, I. 1996. 'Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon' In Wanner L. (ed.), *Lexical Functions in Lexicography and Natural Language Processing, Vol. 1*. Amsterdam: John Benjamins Publishing Company, 37–102. Accessed on 18 September, 2017. http://dialnet.unirioja.es/servlet/extart?codigo=4786249.

Mel'čuk, I. 2006. 'Explanatory Combinatorical Dictionary' In Sica, G. (ed.), *Open Problems in Linguistics and Lexicography*. Monza: Polimetrica International Scientific Publisher, 225–355.

Orešković, M., J. Benić and M. Essert. 2016*a*. 'The Network Integrator of Croatian Lexicographical Resources' In Margalitadze, T. and G. Meladze (eds), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, Tbilisi, Georgia, September 6-10, 2016*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, 267–272. Accessed on 4 September 2017. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex 2016/euralex_2016_026_p267.pdf.

Orešković, M., M. Brajnović and M. Essert. 2017. 'A Step towards Machine Recognition of Tropes', *Book of Abstracts of Third International Symposium on Figurative Thought and Language, FTL3, Osijek, Croatia, April 26-28, 2017*. Osijek: Faculty of Humanities and Social Sciences, University of Osijek, 71.

Orešković, M., M. Čubrilo and M. Essert. 2016*b*. 'The Development of a Network Thesaurus with Morpho-Semantic Word Markups' In Margalitadze, T. and G. Meladze (eds), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, Tbilisi, Georgia, September 6-10, 2016*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, 273–279. Accessed on 4 September, 2017. http://www.euralex.org/elx_proceedings/Euralex2016/euralex_2016_027_p273.pdf.

Orešković, M., J. Topić and M. Essert. 2016*c*. 'Croatian Linguistic System Modules Overview' In Margalitadze, T. and G. Meladze (eds), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, Tbilisi, Georgia, September 6-10, 2016*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, 280–283. Accessed on 4 September, 2017. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex2016/euralex_2016_028_p280.pdf

Polguère, A. 2014. 'From Writing Dictionaries to Weaving Lexical Networks.' *International Journal of Lexicography* 27.4: 396–418. Accessed on 4 September 2017. https://academic.oup.com/ijl/article-abstract/27/4/396/932994.

Prévot, L., C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci and A. Oltramari. 2010. 'Ontology and the Lexicon: A Multi-Disciplinary Perspective' In Huang, C.-R., N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari and L. Prévot (eds), *Ontology and the Lexicon: A Natural Language Processing Perspective*. New York: Cambridge University Press, 3–24.

Pustejovsky, J. 1991. 'The Generative Lexicon.' *Computational linguistics* 17.4: 409–441.

Smith, B. and C. Welty. 2001. 'Ontology: Towards a New Synthesis' In Guarino, N., B. Smith and C. Welty (eds), *Proceedings of the international conference on Formal Ontology in Information Systems – Volume 2001, FOIS '01*, Ogunquit, Maine, USA, October 17-19, 2001. New York: ACM, 3–9. Accessed on 10 October, 2017. https://pdfs.semanticscholar.org/5c2f/16410af0673923dfff874b64fe98f3a75a5b.pdf

Štrkalj Despot, K. and C. Möhrs. 2015. 'Pogled U E-Leksikografiju.' *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 41.2: 329–353.

Wandl-Vogt, E., K. Mörth and A. Bodomo. 2015. *How to Innovate Lexicography by Means of Research Infrastructures - on The European Examples of DARIAH, CLARIN and COST IS 1305 ENeL*. Accessed on 10 October 2017. http://www.slideshare.net/ewv/how-to-innovate-lexicography-by-means-of-research- infrastructures/.