Building the Dictionary of Football Terminology Through Data-Driven and Ontolex Model

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2024.24.1.2

ABSTRACT: This paper investigates the innovative approach to developing a dictionary of football terminology through the application of linked data. Using traditional approaches of dictionary creation includes sequentially written entries and manual analysis of corpora which can be very time consuming. Our approach uses the techniques of computational linguistics and automatization in order to increase the speed of dictionary creation process. The Digital Dictionary of Football Terminology which we have developed is meant for both human and software application use, giving users a fast approach. Furthermore, this gives the possibility of constant updating as well as the integration with different digital resources. This paper focuses on the Serbian Language section of the Serbian-Spanish Dictionary of Football Terminology which is developed within the doctoral dissertation of Jelena Lazarević.

KEYWORDS: dictionary, dictionary of football terminology, corpus, lexical units, computational linguistics.

PAPER SUBMITTED: 10 June 2024
PAPER ACCEPTED: 27 June 2024

Jelena Lazarević
jelazarevic1@gmail.com
ORCID: 0009-0004-4481-2729
PhD student,
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Olivera Kitanović
olivera.kitanovic@rgf.bg.ac.rs
ORCID: 0000-0002-7571-2729
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

The integration of language resources represents one of the principal topics in contemporary research, relating language technologies. There is an increase in the implementation of the data-driven approach for its improved possibilities of data processing possibilities.

A data-driven approach is a methodology leaning on the analysis and interpretation of data as a basis of decision-making and problem-solving. Instead of using intuition or experience through previous theoretical knowledge, a data-driven approach uses concrete, objective and measurable information gathered from the data. This approach is very popular in disciplines such as linguistics, medicine and various information technologies (Kitanović et al. 2021).

This paper will feature researching the possibility of streamlining and improving the traditional approach of creating dictionaries through the example of developing a bilingual dictionary of football terminology.

The traditional approach of creating dictionary entries, one by one, most often in an alphabetic order relies on the analysis within the corpora or excerpted examples with the consultation of referent dictionaries for each individual entry. Using computational linguistics technologies makes this process quicker than the aforementioned practices. Moreover, the entries are not processed sequentially. Instead, the data-driven approach is implemented, led by data and parallel processing of a greater amount of data at once.

Bergh and Ohlander (Bergh and Ohlander 2019) had noticed that XX century features a rise in the use of football terminology in general speech, while some terms had become common in the speech of fans. The language of football is in constant change, adapting itself to events in and around the game. Due to its importance and a great media presence, football, as the People's game, influences general English dictionaries to include football terms as part of the General language.

Digital dictionaries have vast advantages in comparison to traditional, printed dictionaries. According to Fuentes-Olivera and Tarp (Fuertes-Olivera and Tarp 2014), the ideal solution for contemporary dictionaries is in fact their online edition. The first and key advantage of an online dictionary is the speed of search which enables fast word and phrase search, while printed dictionaries need to be manually revisited. The second advantage is the possibility of constant updating of the entries, unlike printed dictionaries which can expire in their relevance within a matter of years.

Online dictionaries also allow the adaptation of fonts, letter size and other parameters to fulfill user needs, while containing additional information, such as detailed etymology and pronunciation. The availability of mobile applications and desktop platforms always enables users access to the dictionary. It is very important to add the possibility of search within a single word entry, to quickly find meaning or information on a certain term while reading.

Putting dictionaries onto digital language learning platforms enables progress following through making word lists, notes, flash cards and other helpful tools that streamline the language learning process. Integrating the dictionaries with the text corpora is very significant because users are allowed to learn words and expressions based on concrete examples of their use. Unlike the global practice, Serbia still features greater commonality levels of traditional dictionary use due to a feeling of authenticity or a simple habit. Currently, digital dictionaries have become a necessary tool for language learners and those who speak multiple languages. At this point, we need to mention that any digital dictionary can be printed.

The Dictionary of Football Terminology featured in this paper is a digital dictionary intended for both human and machine use, regarding software applications. It represents a part of a broader research and the compilation of the first Digital Dictionary of Football in the Serbian language. The entries are translated to Spanish, within the doctoral dissertation of Jelena Lazarević under the title: "Language Characteristics of the New Media Discourse on Football: a Contrastive Analysis of the Serbian and Spanish Language Corpora". The emphasis stays on the connection between computational linguistics and the implementation of technologies of semantical networks in order to fulfill the needs of individual users, as well as software tools.

The Dictionary is intended for the use of individuals working in sport: athletes, coaches, referees, professors, students, sports journalists as well as anyone wishing to master the precise interpretation of footballing terms. This paper particularly focuses on the Serbian section of the bilingual Serbian/Spanish Dictionary of Football Terms.

In Section 2 of this paper we will view the two ways of excerpting information necessary for the creation of the Dictionary: automatic excerption from the corpus *srFudKo*, term candidates, frequencies and examples of use (Krstev et al. 2015; Ivanović et al. 2022), as well as the extraction from the preexisting dictionaries and glossaries available online (Kitanović et al. 2021).

2 Data preparation for the Dictionary compilation

The corpus *srFudKo* (Lazarević et al. 2023) was created from media articles about football in the Serbian language, compiled from five online news sites: *B92, Blic, Mondo, Politika* and *Sport Klub*. The articles were automatically

downloaded through various techniques, which was followed by deduplication of articles, eliminating articles shorter than 3000 characters, sentences in other languages and tables that contained only numerical results. After filtering the articles and text debriding the corpus *srFudKo* was created; it contains 10,100,553 tokens, out of which a total of 8,618,426 represents words, while the rest is interpunction. The corpus content was then annotated by word type and lemmatized though taggers for the Serbian language SrpKor4Tagging-TreeTagger (Stanković et al. 2020; Stanković, Škorić, and Šandrih Todorović 2022).

The corpus *srFudKo* was then used for information excerption. The initial result of the automatic excerption was a list of word candidates, followed by frequency of use. In the first phase, single part terms were extracted, evaluated, and tagged, which was followed by the multiple compound terms (Krstev et al. 2015), most frequent syntactical patterns based on the morphological dictionaries of the Serbian language and local grammars (Krstev 2008). Aside from single part terms, their keyness[1] was also calculated, comparing frequencies in *srFudKo*, with the Contemporary Serbian Language Corpus *SrpKor2013* (Утвић 2011). In terms of compound terms, more complex associative measures were used, such as *T-Score* and *CValue* (Vu, Aw, and Zhang 2008). The details of this stage are described in the paper: "Football terminology: compilation and transformation into *OntoLex-Lemon* resource"(Lazarević et al. 2023).

In terms of information extraction from the preexisting dictionaries and glossaries available online, various files of various types and levels of detail were downloaded. Instead of consulting a dictionary for each individual entry, crossing, harmonizing and processing of all sources was planned to streamline and accelerate the creating of the first contemporary dictionary of football, rich in information. Data integration relied on multiple sources, but the main roles belong to: 1) The four language dictionary based on Serbian, English, French and Spanish terms lists (Mihajlović 2003) and 2) *Kicktionary*[2] (Schmidt 2009), semantically based on the frames of *FrameNet*[3] (Fillmore et al. 2003).

1. https://www.sketchengine.eu/documentation/simple-maths/

2. http://www.kicktionary.de/ the multilingual lexicon that covers terms and their definitions related to football matches, tactics, players, and equipment, with the aim of facilitating learning and understanding of football terminology through visualizations and explanations.

3. *FrameNet* is a lexical database that documents semantic frames and how they connect words to their meanings in different contexts.

Frame Semantics is the theory of meaning focused on conteptual structures – frames, which represent the background necessary for understand words and expressions in various contexts. These frames include associated elements and relations activated while using a certain term, allowing a better understanding of their meaning and context of use. Frames and lexical units in English were crossed with the English side of the QUAD dictionary in order to enrich the Serbian side of the dictionary with semantic frames. Considering that no language of Football Terminology in Serbian had descriptive entry definitions and given the lack of machine translating for the sake of a faster compilation of the Dictionary, we used Glossaries of football terminology in English and Spanish with semi-automatic processing. Aside from *Kictionary*, which also contains certain definitions, dictionaries available at the moment of data processing were also used. We need to add that some of these dictionaries have become unavailable on the previously known addresses:

- The Field[4] Glossary with 900 of the most used football terms is organized through illustrated scenarios. The base language is Portuguese, while containing equivalents in English and Spanish, alongside examples of use. Homonymy and Polysemy was implemented through separated dictionary entries and its structure was inspired through the concepts of Semantic Frames
- The UEFA[5] Football Glossary we have used features a total of 8,086 entries: 1,893 in German, 2,306 in English and 1,887 in French.
- Football Glossary *UsingEnglish*[6] contains different domains and has only 74 entries for frequent terms.
- The NINE language dictionary[7] contains 77 terms with translated equivalents in English, German, French, Spanish, Italian, Portuguese, Polish, Russian and Ukrainian.

The Dictionary of Football Terms *FudLe* was created based on the doctoral dissertation of Jelena Lazarević and contains now a total of 2,648 entries with translation equivalents in Serbian and English that were enriched by

4. previously available on http://dicionariofield.com.br/, see https://www.facebook.com/dicionariofield

5. previously available on https://www.uefa.com/insideuefa/dictionary/

6. https://www.usingenglish.com/glossary/football-vocabulary/

7. https://www.uefa.com/MultimediaFiles/Download/competitions/General/01/80/75/52/1807552_DOWNLOAD.pdf

the word types and frequencies of the *srFudKo* corpus. A total of 93 entries feature an added definition.

Within the Serbian partition, there are: 990 single part term, 996 two-part term, 369 featuring three components, 200 featuring four components, while 123 contain five and more components. Connecting with the semantic frames of the dictionary *Kicktionary* was performed on 142 lexical units. After the extraction of the terminological phrases from the *srFudKo* corpus, manual evaluation and tagging domain markers (whether the term refers to sport or distinctively to football), a list of 1,943 terms was extracted that was also given a code of their syntactic group, frequency, and association measure.

3 Compilation and transformation of the Dictionary into the *Ontolex* Model

The use of the ontology lexicon *OntoLex-Lemon*,[8] short for Ontolex (*LExicon Model for ONtologies*) (McCrae et al. 2017) is on the rise in terms of the implementation of lexical resources as a collection of connected data available online. Determinants, whether they pertain to either one word term or multi-word term from the football domain, alongside the results of extraction from the corpus *srFudKo* (frequencies and examples) were enriched with information downloaded and processed from the dictionaries and glossaries, that were then represented by using the *OntoLex* model.

Aside from the example, dictionary entries point out the word type, of each individual entry and potentially their grammatical attributes. Additional information, such as the definition, meaning, conceptual relations regarding other entries within other resources such as knowledge bases, ontology or lexical bases helps users to better understand the term they are searching for. The context of use through various examples extracted from the corpus, alongside the frequencies are there to complete the terminological image.

The first uses of Lemon, as well as the *OntoLex-Lemon* in Serbian, are connected to the creation of the lexical base Leximirka[9] (Stanković et al. 2018) and the translation of the systems of Serbian electronic dictionaries (Krstev 2008) into a relational base (Рујевић 2022).

8. https://www.w3.org/2016/05/ontolex/ and https://jogracia.github.io/ontolex-lexicog/

9. https://leximirka.jerteh.rs/

This paper features the *OntoLex* model for modeling the RDF[10] (eng. Resource Description Framework), which means representing linguistic descriptions of lexical units from the generated dictionary with the amplification of the lexical layer in order to enable their connections and a machine-provided understanding online. The *OntoLex* modules are introduced through the examples of implementation, beginning with the basic *OntoLex* module which defines lexical units, word forms and lexical meanings. The *vartrans*[11] module contains two important classes: Translation and TranslationSet, where the translation represents a liaison between two lexical meanings, connected to the lexical units in various languages. The *LexInfo*[12] type ontology, type value and property value are used as a catalogue of data categories, such as: grammatical gender, number, word type, etc.

Translation categories are represented by pairing with an external catalogue *OEG Translation Categories.*[13] Other frequently used dictionaries, such as: *Dublin Core, DCMI Metadata Terms*,[14] are used for metadata on sources, authorship, version, or license data. For modeling, recommendations are given in the following guidelines: Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0)[15] (Río Gracia et al. 2023) (Río Gracia et al. 2023), as well as: "Guidelines and best practices for LLOD."

According to the aforementioned recommendations, a bilingual dictionary is organized so that individual languages and separated RDF charts are connected through an additional chart which contains connections of translational equivalents. The generated dictionary is formed by four charts, i.e. collections of data:

- Source lexicon, in this case of the Serbian language,
- Target lexicons, in this case Spanish,
- Compilation of equivalents, also known as *TranslationSet.*

The approach of using resources prepared in the previous phases, shown on the left and bottom parts of the image, as well as models used for transforming them into the RDF form, depicted on the right are illustrated in Figure 1. The output is *FudLe* a bilingual dictionary: in Serbian and Spanish, containing entries enriched with various additional information.

10. https://www.w3.org/RDF/
11. https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans
12. https://lexinfo.net/, https://github.com/ontolex/lexinfo
13. http://purl.org/net/translation-categories
14. http://purl.org/dc/elements/1.1/
15. https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

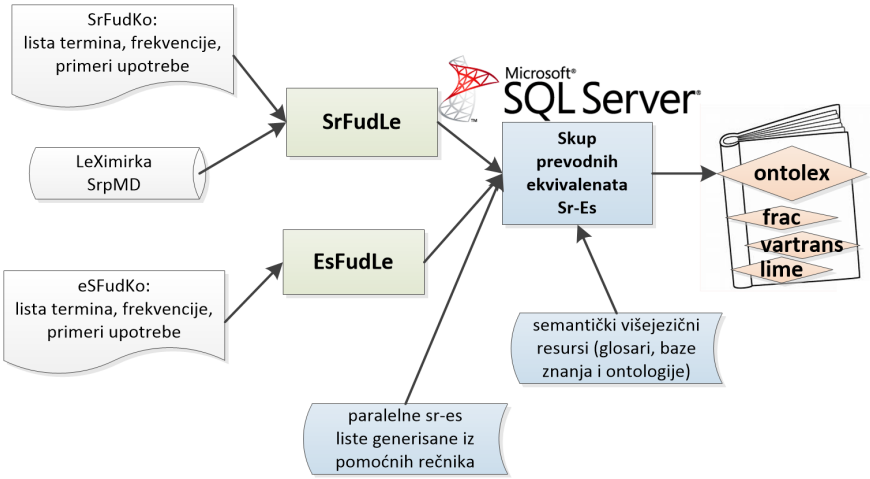

Figure 1. Graphic representation of language data integration from various resources into the *FudLe* Dictionary.

The *OntoLex-FrAC* module (Chiarcos et al. 2020) was used to include information on frequencies, examples in context and collocations extracted from the *srFudKo* corpus into the *FudLe* dictionary. Automatically extracted terms were then manually evaluated, tagged and marked, depending on whether they belonged to the sporting domain or a specialized footballing domain. They were then completed with inflective forms which were further transformed into the *OntoLex* model. The terms were then connected to the examples of use which were chosen by using the GDEX[16] (Good Dictionary Examples) algorithm (Kilgarriff et al. 2008).

The morphological dictionary of compound terms was made by using the tool *LeXimir* (Stanković et al. 2011), and then transformed through the application which follows *OntoLex* specifications and published examples (Chiarcos et al. 2022).

Grammatical information, morpho-syntactic properties of word types were shown in concordance with *LexInfo* vocabulary. The following sections show the use of the basic *OntoLex* Module, as well as the frequency module *OntoLex-FrAC*, examples of use and other corpus-based information (Chiarcos et al. 2022).

16. https://www.sketchengine.eu/documentation/manual-for-gdex/

3.1 The core of the *FudLe*

The corpus *srFudKo* was compiled in Latin script, which means that the proveniente dictionary *FudLe* was also written by using Latin script. The example of the term: *football match* was shown in previous paper (Lazarević et al. 2023) while this paper features the example of the term: yellow card through the *OntoLex* model, which was amplified with additional information on the meaning and the relations towards other resources and languages.

A yellow card is defined as a warning that a referee is giving a player due to their unsportsmanlike conduct. By using the yellow card, the referee is sanctioning a foul in cases where the severity of the conduct hasn't yet merited a red card and immediate expulsion, such as a fight or an intentional foul meant to injure the opponent. This term also exists in the Serbian Morphological Dictionary SrpMD of compound words (Krstev and Vitas 2009) in the form of DELAC script (Savary, Krstev, and Vitas 2007). The source script is the following:

```
žuti(žut.A8:adms1g) karton(karton.N1:ms1q),
NC_AXN+DOM=Sport+Comp+Conc
```

The final transducer NC_AXN generates inflective forms for electronic morphological dictionaries of compound words, where NC marks a noun collocation, and AXN shows a collocation in the form: Adjective-Noun, with the adjective concurring with the noun's grammatical number, gender, case, and animacy. For the components of a compound the transducer provides information on their lemmas, in this case *yellow* and *card*, their corresponding inflection transducers (A8 and N1) and values of the grammatical properties of forms in which they occur (adms1g и ms1q). The grammatical properties are as follows: a-positive grade, d-determined (long) form, m-male grammatical gender, s-singular, 1-nominative case, g-the information on animacy is not important, q-inanimate noun (objects). Most of the grammatical properties are easily mirrored into the ontology *LexInfo*, only the information that the animacy is of no importance is not foreseen. Taking into account the fact that the collocation *yellow card* can be found in various terminological dictionaries, by using the *OntoLex* model, this lexicalized collocation can be tagged with its particular meaning.

A part of the data base for the system *LeXimirka*, implemented within the MS SQL Server (Stanković et al. 2018; Pујевић 2022), is shown in Figure 2 which illustrates tables: *LexicalEntry* and flective forms (marked as *Forms*), as well as compound (marked as Components).

One of the roles of the *LeXimirka* system is to generate flective forms and their grammatical information for single and multi-word terms in the *FudLe* Dictionary. Grammatical information is connected to the inflective forms through Data Categories and their values (*DatCatValues*). The system is equipped with metadata that refer back to the connected information between data categories in Serbian morphological dictionaries and the *LexInfo Dictionary*. A single lexical unit may have various meanings, saved in the table *LexicalSense* which compiles individual categories through the function *SenseProperties* from the *DataCategories* catalogue.

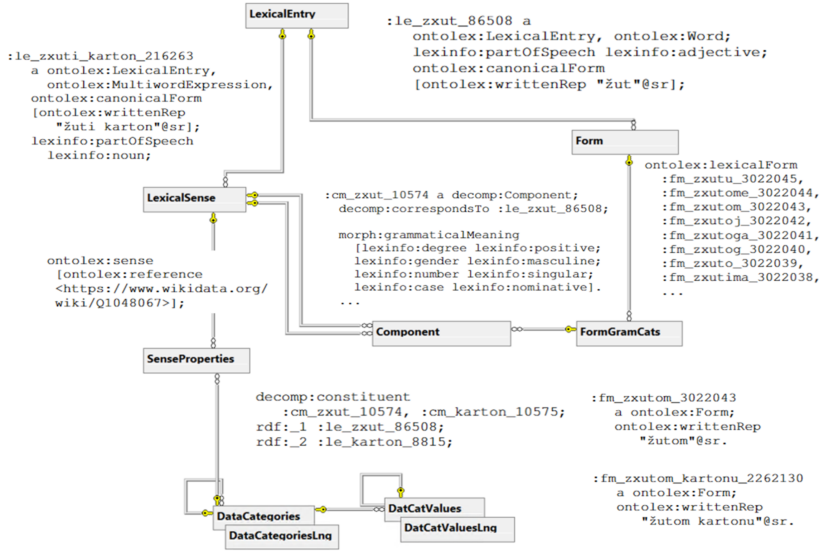

Figure 2. MS SQL Server Data Base diagram with the connected tables and the example of data exported into *OntoLex*.

The script in Figure 2 depicts the collocation *yellow card* with the title of the lexical unit *le_ zxuti_ karton_ 216263*. It consists of the prefix *le*, derived from *LexicalEntry*, the term itself, in this case *yellow card*, and the primary key 216263 in the *LexicalEntry* table within the entire *LeXimirka* data base.

Regarding individual components, the prefix is chosen in a similar way cm, as to mark entries coming from the Component table, with the prefix

fm marking entries from the Form table. The specific letters are changed based on the Aurora coding (Vitas 1979), created by professor Duško Vitas to provide equality in meaning ASCII coding, used for names TITLES whole literals. This means that written forms are coded by using the coding scheme UTF-8.

We can see that the lexical unit *:le_zxuti_karton_216263* belongs to the class *ontolex:LexicalEntry* and that it is a compound component *ontolex:MultiwordExpression*, which is firstly stated in its canonic form, as a lemma *ontolex:canonicalForm*. This is followed by the written form of the lemma, with the added data that in this case is a noun. In terms of meaning, two external resources are used. The first is Wikidata, as a knowledge base, which contains the entry *yellow card* marked with a code *Q1048067*, with the complete URL address as follows: https://www.wikidata.org/wiki/Q1048067. It contains translational equivalents in 38 languages, which means that aside from lexicalization in Serbian, there is also a Spanish translation: *tarjeta amarilla*.

```
:le_zxuti_karton_216263 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "žuti karton"@sr];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>.

  decomp:constituent :cm_zxuti_10574, :cm_karton_10575;
    rdf:_1 :le_zxut_86508;
    rdf:_2 :le_karton_8815.

# komponente kanonskog oblika
:cm_zxuti_10574 a decomp:Component;
  decomp:correspondsTo :le_zxut_86508;

  morph:grammaticalMeaning
    [lexinfo:degree lexinfo:positive;
    lexinfo:gender lexinfo:masculine;
    lexinfo:number lexinfo:singular;
    lexinfo:case lexinfo:nominative].
  ...
```

The corresponding section in Spanish looks as follows:

```
:le_tarjeta_amarilla_10001 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "tarjeta amarilla"@es];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>;

  decomp:constituent :cm_tarjeta_1, :cm_amarilla_2;
    rdf:_1 :le_tarjeta_1;
    rdf:_2 :le_amarillo_2.
```

3.2 Defining meaning in the *FudLe* Dictionary

The class *ontolex:LexicalSense* defines the meaning of a lexical unit when it is interpreted in relation to a pertaining ontological element. The lexical meaning represents a reification of the lexical units being paired with an ontological whole to which it relates. The link between a lexical unit and its ontological entity over the object: *ontolex:LexicalSense* implies that a lexical unit may be used for pairing a given ontological entity.

The idea of a more precise connecting of the lexicon *FrameNet* and the *OntoLex* model has already been shown (Robin, Kulkarni, and Buitelaar 2023). When defining lexical units, it has been shown that they can be used for marking a determined ontological predicate by using the property *ontolex:denotes*. This means that a lexical unit is marking an object class or an ontological element. Also, it is possible to present an evocation of a certain mental concept from the given lexical unit, which doesn't pertain to the class of formal interpretation in a given model.

The class Lexical Concept (*ontolex:LexicalConcept*) represents a mental abstraction, concept or thought unit that can be lexicalized in a way that the class *ontolex:LexicalConcept* represents a subclass of the element *skos:Concept*. The definitions are added to the lexical concepts as a glossa by using property *skos:definition*.

Researching the creation of the semantic football dictionary, while focusing on the football domain corpus, Wu and Li (Wu and Li 2017) have used Word2Vec (Mikolov et al. 2013) to train a vector model for words within the corpus. Combining a statistical algorithm TF-IDF and word vectors, they have processed concepts characteristic for the football domain.

For the football term list, they have calculated the distance from the words within a vector model and separated three of the most proximal terms with the highest cosines value as synonyms for the headword to build a semantic dictionary. They have shown that preexisting semantic dictionaries such as WordNet don't correspond to the needs of certain domains, since the semantic granularity of the words is lower.

It has been necessary to create a semantic dictionary that connects the synonyms to a term, its subordinate and supraordinate terms and definitions in various languages. This is how a semantic dictionary is being created which implements semantic queries of articles from the football domain.

We have created our individual identifiers of concepts based on the gathered resources. The UEFA dictionary terms and definitions was taken as a model for defining the concepts. In the *yellow card* example we see that *:ct_ yellow_ card_ 1* starts with *ct* (abbreviation for Concept), followed by

the term "yellow card" in English, ending in an identifier (for the sake of unambiguity), using an underscore "_" instead of a space.

```
:conceptLexicon a ontolex:ConceptSet .
:le_zxuti_karton_216263 ontolex:evokes :ct_yellow_card_1.
:le_tarjeta_amarilla_10001 ontolex:evokes :ct_yellow_card_1.

:ct_yellow_card_1 a ontolex:LexicalConcept ;
  ontolex:LexicalisedSense :sn_zxuti_karton_216263, :sn_tarjeta_amarilla_10001;
  ontolex:isConceptOf  <https://www.wikidata.org/wiki/Q1048067> ;
  skos:definition "Disciplinska sankcija koju sudija izriče tokom utakmice
    protiv igrača zbog kršenja Pravila igre, posebno zbog lošeg ponašanja u
    sportu ili odlaganja ponovnog početka igre."@sr;
  skos:definición "Sanción disciplinaria impuesta durante un partido por el
    árbitro a un jugador por infracción de las Reglas de Juego, en particular
    por mala conducta deportiva o por retrasar la reanudación del juego."@es;

ontolex:isEvokedBy :le_zxuti_karton_216263, sn_tarjeta_amarilla_10001;
  skos:inScheme :FudLe .
```

Considering that a glossa can be connected to *ontolex:LexicalSense*, which then represents the specificities of use through the property *ontolex:usage*, the interpretation of a lexical unit in relation to the meaning defined within a given ontology is often adapted to the conditions of use or pragmatic implications. Especially due to the registry, connotations, or variations in meaning. These facets can be specified through the use of properties which enables gathering information on the conditions of use and pragmatic implications under which the lexical entry is being referred to a lexical meaning.

The aforementioned conditions of use are not being introduced as a replacement for the formal meaning, but as a complement which better describes a lexical unit. In the case of the *FudLe*, definitions compiled from various sources were extracted from the corpora or directly written in this way.

```
:sn_zxuti_karton_216263 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html>;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_zxuti_karton_216263 ;
  ontolex:usage [ rdf:value "Žuti karton koji sudija vadi iz svog džepa da bi pokazao
    igraču koji je uradio nešto prilično loše, kao što je loš faul ili igranje rukom,
    ali ne tako loš kao prekršaj crvenog kartona kao što je tuča. Žuti karton često
    ide zajedno sa drugom kaznom kao što je slobodan udarac. Dva žuta kartona na jednom
    meču znače crveni karton, a dva žuta u određenom periodu suspenziju."@sr ].

:sn_tarjeta_amarilla_10001 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html> ;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_tarjeta_amarilla_10001 ;
  ontolex:usage [ rdf:value "La tarjeta amarilla se muestra a un jugador por acciones
```

```
como faltas fuertes o tocar el balón con la mano. Puede ir acompañada de sanciones
como tiros libres. Dos tarjetas amarillas en un partido significan una roja, y dos
en un período llevan a suspensión."@es ].
```

3.3 The implementation of the frequencies and their examples in the FudLe dictionary

The documentation for the *OntoLex* module and information on frequencies, examples of use within a given context and other corpora information known as "Frequency, Attestations and Corpus data" (FrAC) are available at the following link: https://github.com/ontolex/frequency-attestation-corpus-information (Chiarcos et al. 2022; Chiarcos et al. 2020).

The *FrAC* module is directed towards the enrichment of dictionaries and other linguistical resources which contain lexicographic data through a model for the representation of statistic data extracted from the corpora: information on frequency and appearance, collocations, markers from lexical resources into corpora and other text collections – the confirmation of their use within the corpora, tagging of corpora and other sources of lexical information, lemmatization according to the dictionary and distributional semantics, meaning vectors of collocations, word, meaning and term embedding, The module treats cases of use within lexicography based on corpus, corporal linguistics, and the processing of natural language in combination with basic and other *OntoLex* modules.

The auxiliary class *:SrFudKo* has been defined in order to provide us with easier handling and shorter annotation. Indications of the frequencies used for *srFudKo* are published on "noSketch" (Kilgarriff et al. 2014) which is maintained by the Society for language resources and technologies – JePTeX.[17]

An auxiliary class *:SrFudKo_lemma_frek* was introduced for the overall frequency of all forms of a concrete term, with a reminder that a term can contain one or various components. The motive for introducing the said class was compacted coding, because it diminishes the number of code segment repetitions. Frequency comparison with the Referent and General Corpus of the Serbian Language СрпКор2021 (Krstev and Stanković 2023; Škorić and Janković 2024) was completed through the noting of frequencies within that corpus by using the object *:SrpKor2021*.

```
# korpus fudbala - srpski
```

17. JePTeX

```
:SrFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrFudKo>] .

:SrFudKo_lemma_freq rdfs:subClassOf
    frac:Frequency, :SrFudKo, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# opšti korpus savremenog srpskog jezika
:SrpKor2021 a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrpKor2021>] .

:SrpKor2021_lemma_freq rdfs:subClassOf
    frac:Frequency, :SrpKor2021, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# korpus fudbala - španski
:EsFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=EsFudKo>] .

:EsFudKo_lemma_freq rdfs:subClassOf
    frac:Frequency, :EsFudKo, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].
```

Shown below is an illustration of the lemmas in the Corpus of Football and the General Corpus of the Serbian Language. One of the factors for determining the terms, also known as *Termhood* is the notion of *Keyness*, which is calculated based on the frequencies in these two corpora. There are two ways of representing frequencies:

```
# frekvencije na nivou leme  (prvi način)
:le_zxut_86508 frac:frequency
    [a :SrFudKo_token_freq; rdf:value "2313" ] .

:le_zxut_86508 frac:frequency
    [a :SrpKor2021_token_freq; rdf:value "23377" ] .

# frekvencije na nivou leme  (drugi način)
:le_zxut_86508 frac:frequency
   :freq_SrFudKo_le_zxut_86508,
   :freq_SrpKor2021_le_zxut_86508.

:freq_SrFudKo_le_zxut_86508
   a :SrFudKo_token_freq; rdf:value "2313".

:freq_SrpKor2021_le_zxut_86508
   a :SrpKor2021_token_freq;
   rdf:value "23377".

# frekvencije na nivou leme za višečlanu reč
:le_zxuti_karton_216263 frac:frequency
```

```
    :freq_SrFudKo_le_zxuti_karton_216263,
    :freq_SrpKor2021_le_zxuti_karton_216263.
```

For the Spanish language, we use the example of frequencies on the *esFudKo* corpus. For the CQL (Corpus Query Language) the query [lemma = "amarillo"], we get a total of 1517 repetitions, out of which the following frequencies are shown in individual forms: amarilla (839), amarillas (244), amarillo (221), amarillos (85), Amarillas (55), Amarilla (41), AMARILLAS (21), Amarillo (10) и AMARILLA (1). Within the FudLe dictionary, frequencies are shown comprised, regardless of caps, which means that amarillas contains 244+55+21=320.

```
# frekvencije na nivou leme (prvi način) španski
:le_amarillo_2 frac:frequency
    [a :EsFudKo_token_freq; rdf:value "1517" ] .

# frekvencije na nivou leme  (drugi način)
:le_amarillo_2 frac:frequency :freq_ESFudKo_le_amarillo_2.

:freq_ErFudKo_le_amarillo_2
   a :EsFudKo_token_freq;
   rdf:value "1517".
```

Absolute frequencies are reached using CQL expressions, while relative frequencies are calculated by a million words – dividing the entire number of words in the corpus, multiplied by a million. In terms of frequencies of complex elements in Serbian and Spanish, one of the possible solutions is shown as follows (Lazarević et al. 2023):

```
SrFudKo_mwe_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].

:EsFudKo_mwe_freq rdfs:subClassOf
  frac:Frequency, :EsFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].
```

The example that follows will show the frequencies of the lemma: “yellow card” (Sr. žuti karton) and the same meaning in Serbian. Using the example of textitsrFudKo, a clarification is given, as well as the queries used to reach their frequencies.

Lemma frequencies are reached through the query “[lemma="žut"][lemma="karton"]”, which yields also the inflected forms: “žuti kartoni” (124), “žuta kartona” (95), “žutih kartona”, “žutog kartona” (84), “žute kartone” (62), “žutim kartonom” (51), “žuti kartoni” (15)...

```
# mwe frekvencija lema

:le_zxuti_karton_216263
  frac:frequency [a :SrFudKo_mwe_freq;  rdf:value "1996"];
  frac:frequency [a :SrpKor2021_mwe_freq;  rdf:value "3192"];
  frac:head :le_karton_8815  .
```

The Spanish Language example is given through the CQL query of the lemma "Tarjeta", meaning "Card": [lemma = "tarjeta"][lemma = "amarillo"], offering 312 hits: "tarjeta amarilla" (214), "tarjetas amarillas" (75), "Tarjetas amarillas" (12), "Tarjeta amarilla" (10), "tarjeta amarillas" (1). When giving frequencies, we count the insensitivity to upper and lowercase.

```
# mwe lemma frequency za primer na španskom

:fm_tarjeta_amarilla_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "224"];
  frac:head :le_tarjeta_1  .

:le_tarjetas_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "87"];
  frac:head :le_tarjeta_1  .

:le_tarjeta_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "1"];
  frac:head :le_tarjeta_1  .
```

It is possible to introduce an individual object for frequencies, used to then connect the data:

```
# frekvencije za španski
:le_tarjeta_amarilla_10001 frac:frequency :freq_le_tarjeta_amarilla_10001.
```

4 Translational equivalents in the *FudLe* dictionary

As previously mentioned, the Dictionary of Football Terminology FudLe comprises of various components, among which we show the source, Serbian section of the Lexicon, the key section in Spanish as well as the Translation Set.

This typology fits more naturally into the base scheme of *OntoLex* and the *Vartans* model. As a result, two separate lexicons in Serbian and Spanish language are published as connected data sets, alongside the translation set which connects them. The publication of the dictionary for other languages is also possible according to the same scheme.

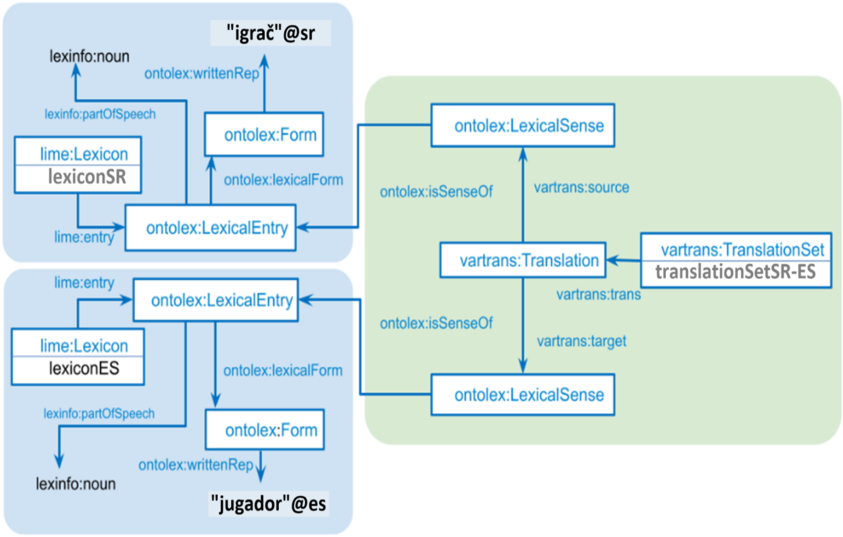

Figure 3. Translation modeling using *OntoLex* and vartrans.

Figure 3 illustrates the presentation scheme used for the translation of the lexical unit "Player" from Serbian into Spanish: "Jugador" within the context of the basic space of *OntoLex* and *Vartans*.

The lexical unit (*ontolex:LexicalEntry*) as well as its attributed properties are used to add lexical information, while *vartrans:Translation* class connects them via lexical meaning *ontolex:LexicalSense*. For example, the property of written form (*ontolex:writtenRep*) connects the subject (*ontolex:Form*) "Player" in Serbian or "Jugador" in Spanish.

There are simpler possibilities in connecting lexical units without defining their circumstantial meaning. In those cases, we use the predicate *vartrans:translatableAs*. However, translational equivalents are established between specific meanings of lexical units and the use of the class *vartrans:Translation* as a conduit between their two meanings.

The following URI identifier was chosen for publishing the dictionary: http://domain/type/concept/reference, with type declaring the type of identified resource. For example: 'id' or 'item' for palpable objects, or 'doc' for the documents describing them; 'def' is used for concepts, while 'set' is used for data sets or elements specific for the context such as: 'authority'

or 'dcterms' (Archer, Goedertier, and Loutas 2012). A similar approach was used for publishing the Serbian-German dictionary SrpNemLex (Andonovski 2023).

In the case of *FudLe*, three data sets were projected:

- The Source data set in Serbian: https://llod.jerteh.rs/id/fudle/lexiconSR
- Key Language data set in Spanish: https://llod.jerteh.rs/id/fudle/lexiconES
- Translational equivalent data set (Spanish-Serbian): https://llod.jerteh.rs/id/fudle/tranSetSR-ES

The example that follows will show the units "igrač"@sr and "jugador"@es with object names adjusted towards easier understanding of the created connections between translational equivalents on the level of word meaning.

```
# :lexiconSR a lime:Lexicon .
...
:lexiconSR lime:entry :lexiconSR/igracy-n-sr .
:lexiconSR/igracy-n-sr a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconSR/igracy-n-sr-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconSR/igracy-n-sr-form a ontolex:Form ;
  ontolex:writtenRep "igrač"@sr .
:lexiconES a lime:Lexicon  .
...

:lexiconES lime:entry :lexiconES/jugador-n-es .
:lexiconES/jugador-n-es a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconES/jugador-n-es-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconES/jugador-n-es-form a ontolex:Form ;
  ontolex:writtenRep "jugador"@es .
...

:tranSetSR-ES a vartrans:TranslationSet   ;
...
:tranSetSR-ES vartrans:trans
  :tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans.
:tranSetSR-ES/igracy_jugador-n-sr-sense a ontolex:LexicalSense ;
  ontolex:isSenseOf :lexiconSR/igracy-n-sr .
:tranSetSR-ES/jugador_igracy-n-es-sense a ontolex:LexicalSense ;
  ontolex:isSenseOf :lexiconES/jugador-n-es .
:tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans
  a vartrans:Translation ;
  vartrans:source :tranSetSR-ES/igracy_jugador-n-sr-sense ;
  vartrans:target :tranSetSR-ES/jugador_igracy-n-es-sense .
```

The module FrAC enables illustrating collocations and embeddings (Chiarcos et al. 2022; Chiarcos et al. 2020), and various vector representations.

5 Conclusion

The development of dictionaries through the methodology of connected data has shown significant advantages in comparison to the traditional approach. Process automatization, use of corpora and integration of various digital resources has allowed a faster compilation of dictionaries. Digital dictionaries offer various advantages, such as: high-speed search, constant updates and user-friendly adaptivity, making them still irreplaceable. The application of techniques provided by computational linguistics and semantical networks has provided the possibility of creating a dictionary which can be used by humans, as well as machines. We expect this approach to be implemented towards creating dictionaries in other fields, which only further enriches the area of lexicography and language technology.

Acknowledgment

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

References

Andonovski, Jelena. 2023. "SrpNemKor and Linked Open Data." *Infotheca - Journal for Digital Humanities* 23 (1): 33–60. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2023.23.1.2.

Archer, Ph., S. Goedertier, and N. Loutas. 2012. *Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC.* Technical report. European Commission.

Bergh, Gunnar, and Sölve Ohlander. 2019. "A hundred years of football English: A dictionary study on the relationship of a special language to general language."

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

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Fuertes-Olivera, Pedro A, and Sven Tarp. 2014. *Theory and practice of specialised online dictionaries: Lexicography versus terminography.* Vol. 146. Walter de Gruyter GmbH & Co KG.

Ivanović, Tanja, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. "Corpus-based bilingual terminology extraction in the power engineering domain." *Terminology* 28 (2): 228–263.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. "GDEX: Automatically finding good dictionary examples in a corpus." In *Proceedings of the XIII EURALEX international congress,* 1:425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra ...

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A data driven approach for raw material terminology." *Applied Sciences* 11 (7): 2892.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

Krstev, Cvetana, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. "Terminology Acquisition and Description Using Lexical Resources and Local Grammars." In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence.* Granada, Spain: LexiCon (Universidad de Granada).

Krstev, Cvetana, and Duško Vitas. 2009. "An Effective Methode for Developing a Comprehensive Morphological E-Dicitonary of Compounds." In *Proceedings of the 28th Conference on Lexis and Grammar,* edited by B. Lamiroy, E. Laporte, and T. Kyriakopoulou, 204–212. 29th September - 3rd October 2009, in Arena Romanistica. Bergen: University of Bergen, Department of Foreign Languages.

Lazarević, Jelena, Ranka Stanković, Mihailo Škorić, and Biljana Rujević. 2023. "Football terminology: compilation and transformation into OntoLex-Lemon resource." In *Proceedings of LDK 2023 – 4th Conference on Language, Data and Knowledge.* 12–15 September, Vienna, Austria. Lisabon: NOVA FCSH - CLUNL. https://doi.org/10.34619/srmk-injj.

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

Mihajlović, Aleksandar. 2003. *Fudbalski rečnik: srpsko-englesko-francusko-španski.* Beograd : A. Mihajlović, 2003 (Beograd : Solidarnost).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26.

Río Gracia, Jorge del, et al. 2023. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0) Draft Community Group Report.* Technical report. 03 October 2023. October. https://bpmlod.github.io/Bilingual_Dictionaries_Report/.

Robin, Cécile, Atharva Kulkarni, and Paul Buitelaar. 2023. "Identifying FrameNet Lexical Semantic Structures for Knowledge Graph Extraction from Financial Customer Interactions." In *Proceedings of the 12th Global Wordnet Conference,* edited by German Rigau, Francis Bond, and Alexandre Rademaker, 91–100. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, January. https://aclanthology.org/2023.gwc-1.11/.

Savary, Agata, Cvetana Krstev, and Duško Vitas. 2007. "Inflectional non compositionality and variation of compounds in French, Polish and Serbian, and their automatic processing." In *Bulag — Bulletin de Linguistique Appliquée et Générale: Les langues slaves et le français : approches formelles dans les études contrastives,* edited by Aleksandra Dziadkiewicz and Izabella Thomas, 73–94. 32. Besançon: Presses Universitaires de Franche Comté.

Schmidt, Thomas. 2009. "The Kicktionary–A multilingual lexical resource of football language." *Multilingual FrameNets in computational lexicography. Methods and applications,* 101–134.

Škorić, Mihailo, and Nikola Janković. 2024. "New Textual Corpora for Serbian Language Modeling." Faculty of Philology, University of Belgrade, *Infotheca* 24 (1).

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries - from File System to lemon Based Lexical Database." In *Proceedings of the 11th International Conference on Language Resources and Evaluation – W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018).* LREC 2018, May 7–12, 2018. Miyazaki, Japan: ELRA.

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference,* 77–84.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of the 12th Language Resources and Evaluation Conference,* 3954–3962. Marseille, France: ELRA.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel bidirectionally pretrained taggers as feature generators." *Applied Sciences* 12 (10): 5028.

Vitas, Duško. 1979. "Prikaz jednog sistema za automatsku obradu teksta." *Zbornik radova XIV jugoslovenskog medunarodnog simpozijuma o obradi podataka Informatica* 79:7–101.

Vu, Thuy, Aiti Aw, and Min Zhang. 2008. "Term extraction through unithood and termhood unification." In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.*

Wu, Jiguang, and Ying Li. 2017. "Research on construction of semantic dictionary in the football field." In *Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA),* 303–306. IEEE Xplore Full-Text PDF available. IEEE.

Рујевић, Биљана. 2022. "Речници у дигиталном добу – информатичка подршка за српски језик." Докторска дисертација, Универзитет у Београду, Филолошки факултет.

Утвић, Милош. 2011. "Анотација Корпуса савременог српског језика." *ИНФОтека* 12, no. 2 (December): 39–51.