Football terminology: compilation and transformation into OntoLex-Lemon resource

Jelena Lazarević
University of Belgrade Faculty of Philology
Serbia
jelazarevic1@gmail.com

Ranka Stanković and **Mihailo Škorić** and **Biljana Rujević**
University of Belgrade
Faculty of Mining and Geology
Serbia
(ranka.stankovic|mihailo.skoric|biljana.rujevic)@rgf.bg.ac.rs

Abstract

The purpose of this article is to present the ongoing project which is the compilation of the first digital Football dictionary in the Serbian language, as well as to demonstrate the application of OntoLex and associated modules. The OntoLex-FrAC module for a football-specific dictionary includes information about frequency, attestation, and corpus usage. In this case, a domain-specific corpus was created by the name of SrFudKo, containing news articles about football in Serbian. Multi-word terms were automatically extracted from the Serbian corpus, then manually evaluated and classified as either sport or football-related. An inflection lexicon was produced and transformed into the OntoLex-Lemon format, Frequency information from the extraction phase was assigned to the entries. Finally, a few lexical entries were linked with the attestations from the corpus.

1 Introduction

This paper will use the expression "the language of football", as a reference to the terminology and specialized expressions used relating to football. We are aware that this is not a language in a traditional sense, but rather a specific type of jargon belonging to the domain of sporting terminology. Said terminology includes terms such as goal, corner, throw-in, offside, etc. These concepts are essential for understanding and communication about football. Here are some of the terms used, related to football:[1]

- Goal: *fundamental scoring event in football, that occurs when a player successfully kicks the ball crosses the goal line of the opposing team, typically resulting in one point being awarded to the team that scored.*
- Corner: *restart of play that occurs when the attacking team plays the ball from the corner of the field towards the opposing team's goal.*
- Offside: *position where a player plays the ball more advanced than the position of the last player on the opposing team.*
- Foul: *break of the rules of the game by a player making contact with an opponent.*

In addition to the use of said terminology and specialized expressions, there is a specific writing style, which emphasizes important events and moments during the match. Moreover, journalists often use technical terms and match analysis, in order to explain tactical decisions and player performances. They also rely on statistical data and analysis for the sake of adding depth and context to their reports.

To recapitulate, news articles about football use a specific language, crafted to provide accurate and informative reports about the sport. Shown here are examples of characteristic multi-worded expressions often used in articles about football:

- Potentially dangerous situation: *situation where the ball is near the goal and there is a strong possibility that the opposing team may score.*
- Effective play: *team's use of sound tactics and strategies, resulting in positive outcomes.*
- Strong play: *style of play in which a team employs physicality, often utilizing powerful kicks and high jumps, to gain an advantage over their opponents.*
- Best chance: *situation in which a player has a favorable opportunity to score a goal, often resulting from a good position or a well-placed pass.*

[1]The definitions are adapted from UEFA dictionary https://www.uefa.com/insideuefa/dictionary/ and Wikipedia's Glossary of association football terms https://en.wikipedia.org/wiki/Glossary_of_association_football_terms

There are currently no digital terminological dictionaries that cover this area in the Serbian language, which is the main motivation for creating a Serbian lexicon of football terms and expressions. There is a traditional, analog dictionary (Mihajlović, 2003) that covers four languages: Serbian, English, French, and Spanish. It is mentioned in the review of Serbian-Spanish dictionaries (Pejovic, 2021), but its structure does not meet the requirements of contemporary lexicography: lexical entry contains only translation equivalents. The number of terms and their selection are subjective, while the development of the dictionary was not corpus-driven.

De Oliveira Chishman et al. (2015) discussed the relevance of the Sketch Engine software (Kilgarriff et al., 2014) to build Field–Football Expressions Dictionary[2], a trilingual terminological resource based on the notion of the frame and on linguistic corpora. They described the analysis procedures to identify polysemic words and collocations in the corpus. Its building process involved, amongst other stages, the compilation of three comparable corpora for Spanish, Portuguese, and English.

Bergh and Ohlander (2019) have shown that, over the past hundred years, football vocabulary has become more *mainstream*, while some football terms have formed a strong presence in the minds of fans. Thus, the language of football remains in a state of constant flux, responding to the developments in and around the game. They conclude that due to its status and large media coverage of the "people's game", the English general purpose dictionaries are recognizing more of this footballing vocabulary as part of the general language.

The language of sport has always been a field of rich specialized linguistic communication (Liponski, 2009). Within sports, football is an especially important element of communication (Penn, 2016), due to the fact that in general human communication, football represents a significant topic. Communication about sports is primarily carried out by the media in constant contact with their target group of readers – sports fans. The language of sports, especially in Europe, is primarily the language of football, which has therefore turned into a public discourse accessible to all (Bergh and Ohlander, 2012).

The language of sports and therefore of sports journalism differs from other forms of expression. Compared to literary language, there are differences in the degree of formality of expression and the style of presenting information. The use of collocations and idioms is present in the media coverage, which makes the articles seem much closer to the readers.

In his research, Čudomirović (2014) analyzed how the media constructed the national identity of the Serbian National Team during the 2010 World Cup matches. The corpus used for analysis included 35 reports from daily newspapers in Serbia. His findings showed that the press constructed the Serbian national identity as both highly homogeneous and self-focused, with an emphasis on achieving and maintaining unity within the nation.

There are numerous examples of research in the field of football language worldwide. However, the most interesting is *Kicktionary*, a multilingual (German – English – French) electronic dictionary of the football language, that includes 1926 football terms, of which 599 are in English, 792 in German and 535 in French (Schmidt, 2009). The terms are structured into a hierarchy of scenarios and frameworks, which further include multiple concepts. Each word is illustrated with one or more example sentences from the authentic: written or spoken football language.

The main goal of the *Kicktionary* project was to explore how the linguistic theories of lexical semantics, as well as corpus linguistic methods, hypertext technologies, and computational language-processing techniques, can help to create a lexical resource – better than, or at least different from, traditional analog dictionaries. Although primarily intended for humans, *Kicktionary* has also been used to create models for automatic text markup. Specifically, an adapted version of the frame semantic parsing model *LOME* was used to automatically label texts with frames and semantic roles according to the *Kicktionary* lexical resource (Minnema, 2021).

Inspired by numerous works, our research question is the following: Is it possible to semi-automatically generate a list of terms and football expressions for the Serbian language?

The Section 2 Materials and methods will firstly present the dataset, i.e., the corpus of texts used for the research, the usage and dictionary microstructure, the methods of automatic extraction of terms and manual evaluation criteria, followed by a short

[2]http://dicionariofield.com.br/langselect

outline of the OntoLex-Lemon[3] core model (McCrae et al., 2017), a widely used vocabulary for modeling machine-readable dictionaries on the Semantic Web and as Linguistic Linked Open Data (LLOD), with extension Morph[4] (Klimek et al., 2019; Chiarcos et al., 2022c) and OntoLex-FrAC module (Chiarcos et al., 2022a, 2020).

The Section 3 is dedicated to the results, where the typical examples for several observed syntactic groups will be shown. The Section 4 is dedicated to the examples of lexical entries published in the form of linked data following the OntoLex-Lemon and OntoLex-FrAC specifications. Ultimately, this study offers conclusive considerations and directions for further research.

2 Materials and Methods

2.1 FudKo Corpus

The *srFudKo* corpus is comprised of articles about football in the Serbian language. These articles are gathered from five Serbian digital news sites: *B92*, *Blic*, *Mondo*, *Politika*, and *Sport Klub*. The articles were automatically retrieved through various web scraping techniques, following the harmonization of the gathered structure, and the text was cleansed. Articles shorter than 3000 characters, sentences in other languages, and tables containing only numerical results were eliminated. The article titles were also analyzed, resulting in the removal of 130 duplicate articles detected by their titles. They were then manually examined and removed.

The corpus was prepared as a collection of *XML* files, in which articles are marked with the following structural labels: `<data>` - the basic elements of each document, `<post>` - published article, `<date>` - article date, `<title>` - article title, and `<p>` - paragraph or text of the article. XML files are organized by year and by the portals from which they were downloaded, so 11,117 articles are distributed across 37 files.

Regarding the distribution across portals, *Politika* is the most represented with 3257 articles. They are followed by *Mondo* with 2639 articles, *B92* with 2514 articles, *Sport klub* with 1937 articles, and *Blic* with 770 articles (Table 1). The articles taken from the *Politika* website cover a long period from 2006 to 2021, making this the largest partition. *Sport Klub* covered the years 2017 to 2021, while *Mondo* covered the years 2013 to 2021. The *B92* website was downloaded from 2017 to 2021, and *Blic* was parsed for only two years: 2020 and 2021, making this partition the smallest.

Portal	Period	Number of Articles	Number of Words
Politika	2006-2021	3257	3.1M
Mondo	2013-2021	2639	2.8M
B92	2013-2021	2514	1.9M
Sport klub	2017-2021	1937	1.6M
Blic	2020-2021	770	0.7M

Table 1: Articles distribution across portals

The corpus was tagged with part-of-speech and lemma using a tagger: *SrpKor4Tagging-TreeTagger* for the Serbian language [5] (Stanković et al., 2020; Stanković et al., 2022) integrated into the *TXM tool* (Heiden, 2010). The tagger was trained on the manually annotated corpus *SrpKor4Tagging*[6], which combines literary one-third and administrative two-thirds texts in Serbian.

The corpus was annotated with two sets of part-of-speech tags: *Universal POS* and *SrpLemKor* (a set created based on the traditional, descriptive grammar of the Serbian language), and lemmatized, containing 342,803 tokens. The lemmatization is based on electronic morphological dictionaries for Serbian (Krstev, 2008; Vitas and Krstev, 2012), specifically on the derived distribution intended for tagging *SrpMD4Tagging*[7] (Serbian Morphological Dictionaries for Tagging).

The TXM platform has proven to be very successful for corpus analysis, frequency distributions, and visual presentation. After filtering articles and cleaning the text, the *srFudKo* corpus contains 10,100,553 tokens, of which 8,618,426 are words, and the remainder consists of punctuation marks.

2.2 Dictionary Usage and Microstructure

A sports dictionary of football can be useful for various individuals involved in the sport. They include players, coaches, referees, commentators, journalists, and fans who wish to enhance their understanding and communication in the realm of

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

[4] https://www.w3.org/community/ontolex/wiki/Morphology

[5] https://live.european-language-grid.eu/catalogue/ld/9296

[6] https://live.european-language-grid.eu/catalogue/corpus/9295

[7] https://live.european-language-grid.eu/catalogue/lcr/9294

football. This dictionary will also be used in NLP (Natural Language Processing) applications related to the football domain.

Football players, both amateur and professional, can benefit from a sport dictionary of football, helping enhance their understanding of technical terms, rules, positions, tactics, and strategies used in the game. Thus it can help them communicate effectively with their teammates and coaches. This is especially helpful in the case of foreign players that do not speak the native language of their teammates. Football coaches can also utilize this type of dictionary to reinforce their knowledge of the game and stay updated in the latest terminology. It can assist them in explaining concepts to players, designing training sessions, and developing game plans.

Referees and officials responsible for enforcing football rules can use this football dictionary to ensure a comprehensive understanding of the terms used in the game. This helps them make accurate decisions and maintain consistency during matches. Commentators and analysts who provide match commentary or analysis can utilize a football dictionary to expand their vocabulary and improve their understanding of the game. It allows them to deliver more informative and engaging commentary to viewers. Football journalists and writers can reference a specialized sporting dictionary of football to ensure accuracy in their match reports, using appropriate terminology while discussing player profiles, match analysis, or tactical elements.

Football fans who wish to deepen their knowledge of the sport can benefit from a football dictionary, which enables them to understand better match broadcasts, articles, discussions, and conversations related to the sport. It also enhances their overall enjoyment and engagement with the game.

The microstructure of this football dictionary will include a range of information related to lemma (base word), inflected forms, examples or attestations, frequencies, multi-word expressions, and collocations. Here's a breakdown of each component:

- The lemma represents the base, canonical form, and serves as the entry point in the dictionary. For example, in the football domain a lemma could be *gol* (*goal*) or *udarac* (*kick*).
- The inflected forms of a lemma are important for Serbian as a highly inflected language. For instance, variations of the lemma *udarac* could include *udarca, udarcu, udarci, udarcima, udarce*, etc.
- The illustrative examples or attestations showcase the usage of the lemma in different contexts. These examples demonstrate how the word is used in football-related sentences or phrases.
- The multiword expressions, including fixed phrases, idioms, or collocations specific to the football domain will be included and related to its component words.
- Word usage frequency indicates how common or uncommon a word is within the football domain. Frequencies will be represented through numerical values, both in domain-specific football corpus and in the general-purpose Corpus of the contemporary Serbian language *SrpKor2013* (Utvić, 2011; Vitas and Krstev, 2012), as illustrated through the examples in the Section 4.
- The term collocation refers to words that frequently occur together with a specific lemma. In a dictionary focused on football, collocations can highlight common word combinations or phrases that involve the main lemma.

The current focus is based on a monolingual dictionary. However, future research will include term translation equivalents in the target language. These would also provide corresponding phrases or idioms in the other language, allowing users to understand football-related expressions in both languages. It is important to state that definitions are not part of the initial phase but are planned for the following phase. This is due to the fact that the initial phase is focused on automatic procedures that are already developed. For the definition extraction in Serbian, initial results are presented in (Stanković et al., 2021). However, the solution requires improvement and adaptation for this particular case of use.

Including multi-word expressions and their bilingual equivalents will enhance the dictionary's coverage of idiomatic and context-specific language usage in the football domain, helping users grasp the nuances and intricacies of the language related to the sport.

The outlined micro-structure of the football dictionary aims to provide comprehensive information

about the lemma, its variations, usage examples, frequency of occurrence, and common word combinations, allowing users to better understand and utilize football-related vocabulary.

2.3 Terminology Extraction Approach

The process of football terminology extraction from the corpus *srFudKo* included the following steps:

1. automatic extraction of candidates,
2. manual evaluation and classification,
3. import to lexical database,
4. export to other formats (DELA[8] for Unitex[9], RDF, etc.).

The statistical measure *Keyness* is used in the step of terminology extraction, for identifying terms that are significantly more frequent in the football corpus *srFudKo*, compared to the Corpus of contemporary Serbian *SrpKor2013* (Utvić, 2011; Vitas and Krstev, 2012). The relevance and specificity of a term within a football domain are calculated through the ratio of term frequency in the corpus *srFudKo*, as the target corpus, compared to its frequency in *SrpKor2013*, as the reference corpus. The terms with a high keyness score are considered to be highly relevant, distinct to the football domain, and thus can be used as potential candidates for the terminology lexicon. The keyness function was applied to single-word lemma and multi-words extracted with various syntactic patterns (Krstev et al., 2015).

Multi-word candidates are extracted from texts in their various inflected forms using lexical resources and local grammars developed for Serbian (Krstev et al., 2015) with patterns explained in Section 3. The lemmatization of extracted multi-word candidates, that is, their linking to one normalized form is of low importance for the English language. However, in terms of highly-inflected languages, such as Serbian and other Slavic languages, this task can hardly be avoided, as each nominal multi-word unit (MWU)[10] can have many inflected forms (from five to ten or even more) and many of these forms (but usually not all) can, in general, be extracted from a corpus (Krstev et al., 2015).

The hybrid system called *Srp-TE* (Stanković et al., 2016) was used, which relies on the application of syntactic patterns and electronic Morphological dictionaries for the Serbian language *SrpMD* (Krstev, 2008) that contain both single and multi-word units, covering general lexicon, proper names, toponyms, encyclopedic knowledge, and terminology from numerous domains.

Class names correspond to FSTs (Finite-state transducers) used for the inflection of MWUs belonging to that class. For example, MWUs are composed of an adjective (A) followed by a noun (N), which concord in gender, number, case, and animacy, belong to the AXN class. The letter X represents a component that remains unchanged when the MWU inflects. It can also denote a separator, like a space or a hyphen. The number preceding X indicates how many of these parts there are in the MWU, with 2X representing two uninflected components, one of which is a separator, 4X representing four components, two of which are separators, etc.

The most frequent syntactic structures, for example AXN, 2XN, N2X, N4X, AXN2X, NXN, AXAXN, N6X, AXN4X, 2XAXN, AXN6X, N8X, are implemented. In the Section 3 explanations are given, with examples for the most productive syntactic structures.

2.4 Ontolex-lemon and OntoLex-FrAC

The use of the OntoLex-Lemon is increasing in terms of lexical resources in the web of data. The lexical entries (single and multi-words) from the football domain, extracted by the approach described in the Section 2 are represented using the OntoLex-Lemon.

The morphological dictionary of multi-word units was produced using a multipurpose tool (Stanković et al., 2011), then transformed with a custom application, following the OntoLex specifications, and published examples (Chiarcos et al., 2022b). The grammatical information, morpho-syntactic features about word forms were given by tag properties in accordance with the *LexInfo vocabulary*[11].

The Section 4.2 presents the use of the OntoLex core module and the module for Frequency, Attes-

[8] Dictionnaires electroniques du LADL - Laboratoire d'Automatique Documentaire et Linguistique

[9] https://unitexgramlab.org

[10] We use the term *multi-word unit* as a general term for MWE, collocation, multi-word term, or multi-word named entity

[11] https://lexinfo.net/

tations, and Corpus-Based Information (OntoLex-FrAC) (Chiarcos et al., 2022a). The information found in the corpora, such as attestations and frequency information of tokens (forms) and lemmas (lexical entries) that are automatically derived from corpora, are introduced following the OntoLex-FrAC specifications.

3 Terminology Extraction Results

The terminology extraction in this research study relies upon the results of previous research, both for building and using a terminology system, which includes data, application, and user-interface layers, covering different data and software technologies. The rule-based automatic multi-word term extraction and lemmatization are first used in the domain of library-information terminology (Krstev et al., 2015; Stanković et al., 2016). This data-driven approach was used for raw material terminology (Kitanović et al., 2021), and corpus-based bilingual terminology extraction in the power engineering domain (Ivanović et al., 2022).

The conversion of electronic dictionaries from a file system to a lexical database *LeXimirka*, based on the Lemon model has resulted in a robust system, that not only manages electronic dictionaries but also incorporates a connection with corpora, including results of systems for automatic - single and multi-word terminology extraction (Stanković et al., 2018; Lazić and Škorić, 2020).

Figure 1 presents a web form with lexical entry *utakmica* ('match, sports competition') several parts:

1. inflected forms with grammatical information,
2. concordances in the selected corpus, in this case *srFudKo*,
3. frequencies of inflected forms in the selected corpus for lexical entry of syntactic patterns, in this case, adjective-noun `A(N)`, where the noun is the current lexical entry,
4. lemma frequencies, where in case of syntactic patterns, all components are lemmatized,
5. links to multi-word lexical entries in *LeXimirka* where current entry is one component.

Before the extraction procedure was conducted as part of this research study, the football domain was not specifically processed. However, the electronic morphological dictionary already had a number of terms related to the sporting domain. Using the marker `DOM=Sport`, a total of 185 simple words and 240 multi-word units were marked, belonging to the domain of sport. The semantic marker `+Sport` denoting sporting disciplines was assigned to four simple words and 13 multi-word units. After processing the football domain corpus *SrFudKo* in the Serbian language, some additional entries were prepared. A new marker `DOM=Fudbal` was introduced for the football domain. The list of candidates already in the morphological dictionary was extracted using the keyness function and a new marker was assigned, based on annotations from two independent evaluators and a supervisor that resolved differences. The first author was one of the evaluators, and she has nearly a decade-long experience in sports journalism, reporting primarily on football and creating football-themed articles in multiple languages, which allows her to offer her practical expertise to the academic realm. The second evaluator is a dedicated enthusiast of football.

As for the nouns, a total of 915 nouns that are characteristic of football and sporting articles were marked, while an additional 219 nouns were marked as belonging to the football domain (e.g. *gol, fudbal, fudbaler, poluvreme, golman, mreža, penal* (*goal, football, football player, halftime, goalkeeper, net, penalty*)). When it comes to verbs, there are 196 sports and 5 specific football terms (e.g. *predriblati, uklizati, uštopovati, proklizati* (*to feint, to tackle, to intercept, to slide tackle*)).

Presented here are some of the most productive patterns:

- AXN – an adjective followed by a noun; the adjective and the noun have to concord in all four grammatical categories; e.g. *bela tačka, crveni karton, fudbalski klub*, (*penalty mark, red card, football club*),
- N2X – a noun followed by a non-inflecting word, usually a noun in the genitive or in the instrumental case; e.g. *OFK Beograd, het-trik, FS Srbija, plej-aut*, (*OFK Belgrade, hat-trick, FS Serbia, play-out*),
- N4X – a noun followed by two words that do not inflect in the MWU: 1) A noun followed by a prepositional phrase; e.g. *uzbuđenje pred golom, centaršut iz kornera*, (*excitement in*

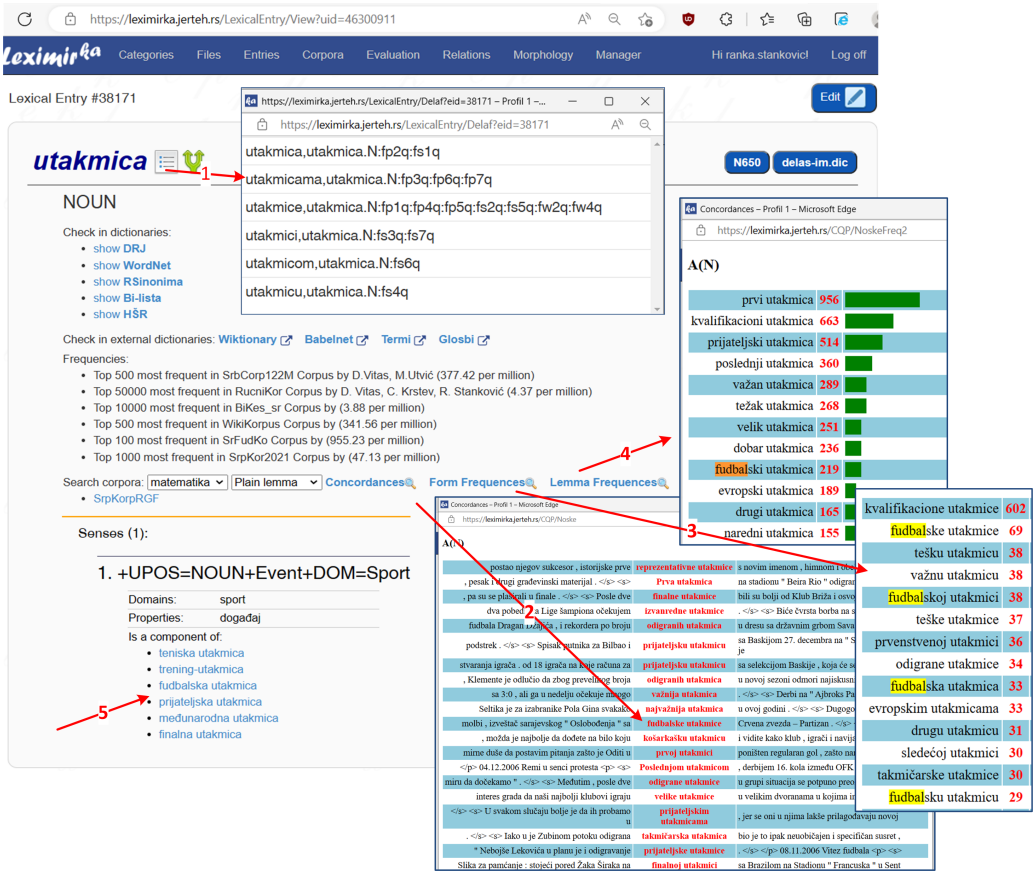

Figure 1: Panel from LeXimirka lexical resources management system

front of the goal, corner kick); 2) A noun followed by two adjectives/nouns in the genitive case or instrumental case; e.g. *ivica kaznenog prostora, utakmica visokog rizika*, (*edge of the penalty area, high-risk match* (a match with potential for violence or disturbances)),

- AXN2X – a noun preceded by an adjective concurring in the gender, number, case and animateness and followed by a word that does not inflect in the MWU, usually a noun in the genitive or instrumental case; e.g. *grupna faza lige, prvo kolo kvalifikacija, evropska kuća fudbala*, (*group stage of the league, first qualifying round, the Union of European Football Associations* (UEFA)) ,

- AXAXN – a noun preceded by two adjectives, concurring in gender, number, case and animateness; e.g. *zimski prelazni rok, Svetsko fudbalsko prvenstvo, centralni vezni igrač*, (*winter transfer window , FIFA World Cup, central midfield player*),

- N6X - a noun followed by three words that do not inflect in the MWU: *učešće u ligi šampiona, udarac sa ivice šesnaesterca, borba na sredini terena*, (*participation in the Champions League, shot from the edge of the penalty area, a battle in the middle of the field*),

- AXN4X – a noun preceded by an adjective concurring in the gender, number, case and animateness, followed by two words that do not inflect in the MWU or by two adjectives/-nouns in the genitive case or in the instrumental case: *Svetsko prvenstvo u fudbalu, prvo mesto na tabeli, žuti karton zbog simuliranja*, (*FIFA World Cup, first place on the table, yellow card for simulation*),

- 2XAXN - an adjective followed by a noun concurring all four grammatical categories, preceded by a word that does not inflect in the MWU; *FK Crvena zvezda, crveno-beli dres, crno-beli tabor*, (*FC Red Star, the Red and White jersey, the Black and White side*),

- N8X - a noun followed by four words that do not inflect in the MWU: *udarac sa ivice kaznenog prostora, bod u borbi za opstanak*,

(*shot from the edge of the penalty area, point in the fight for survival*).

4 FudLe: Linked Data Lexicon

4.1 OntoLex Core Part of FudLe

We illustrate the conversion of electronic dictionary entries with the term *fudbalska utakmica* (eng. football match), which is a competition between two football teams. In Serbian Morphological E-Dictionary (SrpMD) of Compounds (Krstev and Vitas, 2009) in the form of DELAC (Savary et al., 2007) the original dictionary entry is:

```
fudbalska(fudbalski.A2:aefs1g)
utakmica(utakmica.N650:fs1q),
NC_AXN+DOM=Sport+Comp
```

The finite state transducer (FST) `NC_AXN` generates the inflected forms for the morphological e-dictionaries of compound words, where NC stands for Noun compound and AXN depicts adjective-noun compound, where the adjective concurs with the noun in its grammatical number, gender, case, and animacy. For the components that the FST inflects, it requires information about lemma (*fudbalski* and *utakmica*), the FST (`A2` and `N650`) for simple component word and values for grammatical features (aefs1g and fs1q).

The grammatical features are: *a* - positive degree, *e* - form both definite and indefinite, *f* - feminine grammatical gender, *s* - singular number, *1* - nominative case, *g* - no consequence for animacy, *q* - inanimate. Most of the grammatical features are easily mapped to *Lexinfo* but the dilemma for their mapping was `lexinfo:otherAnimacy` adequate for *g* - no consequence for animacy and for the forms that are both definite and indefinite.

Here, we assume that the term *fudbalska utakmica* is a multi-word expression, since it is in the SMD and it can be found in terminological dictionaries. However, it can be treated as a collocate as well. By using the *OntoLex-Lemon* vocabulary, we can declare that it is a (lexicalized) MWU with its specific meaning.

A part of the *LeXimirka* MS SQL Server database's data model, is shown in Figure 2, which displays tables for lexical entries and inflected forms, as well as components for multi-word units. Grammatical information is linked to the inflected forms through data categories and their values. The system is provided with metadata related to linked information between data categories in the Serbian morphological dictionaries and the *Lexinfo* vocabulary.

The following listing presents an example of a multi-word unit, where the name: *le_fudbalska_utakmica_220902* is composed of prefix *le* that stands for *LexicalEntry*, term *fudbalska_utakmica* and primary key *220902* from the table *LexicalEntry* from database *LeXimirka*. Similarly, prefix *cm* denote entries from the table *Component* and prefix *fm* denote entries from the table *Form*.

```
:le_fudbalska_utakmica_220902
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "fudbalska utakmica"@sr];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference <https://
    dbpedia.org/ontology/FootballMatch>];
  decomp:constituent :cm_fudbalska_20258;
  decomp:constituent :cm_utakmica_20259;
  rdf:_1 :le_fudbalski_78369; # lexical
  rdf:_2 :le_utakmica_38171.  # entries

# component of cannonical form
:cm_fudbalska_20258  a decomp:Component;
  decomp:correspondsTo :le_fudbalski_78369;
  morph:grammaticalMeaning
    [lexinfo:degree lexinfo:positive;
    lexinfo:gender lexinfo:feminine;
    lexinfo:number lexinfo:singular;
    lexinfo:case lexinfo:nominative;
    lexinfo: lexinfo:inanimate].
  ...
```

The inflected forms of single and multi-word units in morphological dictionaries are followed by a set of data category values. The majority of inflected forms have ambiguous grammatical interpretations. The following example presents typical instances of single and multi-word units - *fudbalska utakmica*.

```
# inflected forms for adjective
fudbalska:aefs1g:aefs5g:aemw2g:aemw4g...
fudbalskoj:aefs3g:aefs7g
fudbalskim:aefp3g:aefp6g:aefp7g:aemp3g...
...
# inflected forms for noun
utakmica:fp3q:fp6q:fp7q
utakmici:fs3q:fs7q
utakmicama:fs6q
...
# multiword inflected forms
fudbalska utakmica:fs1q
fudbalskoj utakmici:fs3q:fs7q
fudbalskim utakmicama:fp3q:fp6q:fp7q
...
```

The following example presents the first lexical entry - the adjective component *fudbalski* with a

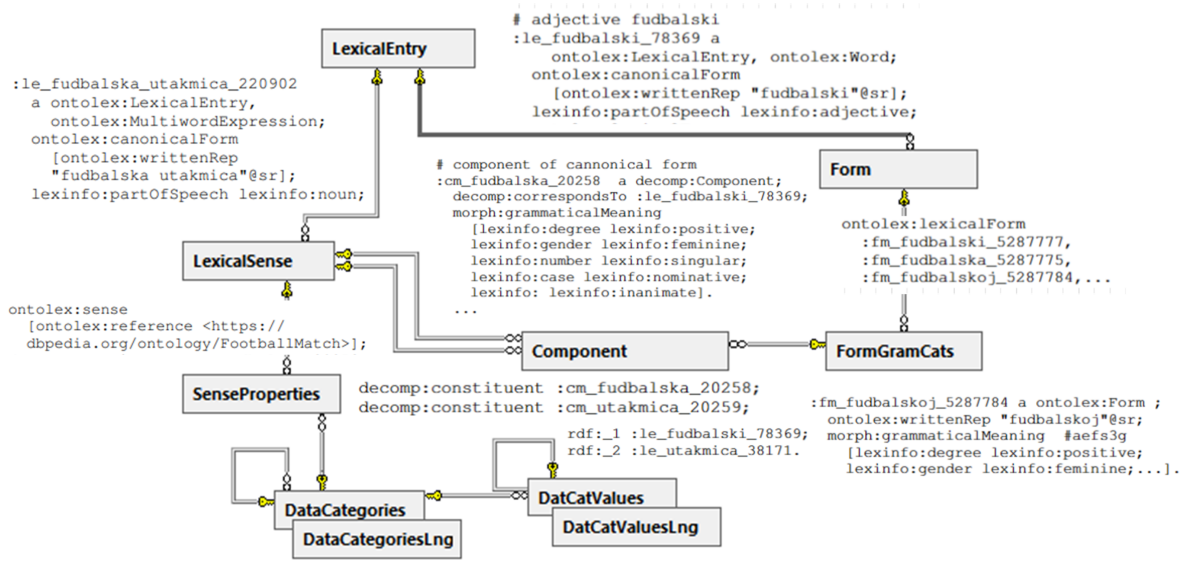

Figure 2: MS SQL Server database diagram with tables related to lexical entries and inflected form

sample of the inflected forms, accompanied by its grammatical information. It is followed by a lexical entry *utakmica* as the second component in its inflected form *utakmici*.

```
# adjective fudbalski
:le_fudbalski_78369 a
    ontolex:LexicalEntry, ontolex:Word;
  ontolex:canonicalForm
    [ontolex:writtenRep "fudbalski"@sr];
  lexinfo:partOfSpeech lexinfo:adjective;
  ontolex:lexicalForm
    :fm_fudbalski_5287777,
    :fm_fudbalska_5287775,
    :fm_fudbalskoj_5287784,...
:fm_fudbalskoj_5287784 a ontolex:Form ;
  ontolex:writtenRep "fudbalskoj"@sr;
  morph:grammaticalMeaning  #aefs3g
    [lexinfo:degree lexinfo:positive;
    lexinfo:gender lexinfo:feminine;...].
 ...
 # noun utakmica
 :le_utakmica_38171 a
    ontolex:LexicalEntry, ontolex:Word;
  ontolex:canonicalForm
    [ontolex:writtenRep "utakmica"@sr];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:lexicalForm
    :fm_utakmica_4569852,
    :fm_utakmice_4569854,
    :fm_utakmici_4569855,...
:fm_utakmici_4569855 a ontolex:Form ;
  ontolex:writtenRep "utakmici"@sr;
  morph:grammaticalMeaning  #fs3q
    [lexinfo:gender lexinfo:feminine;
    lexinfo:number lexinfo:singuar;...].
 ...
```

The examples of inflected forms a in multi-word lexical entry *fudbalska utakmica* and form *fudbalskoj utakmici* is given with its grammatical information:

```
:le_fudbalska_utakmica_220902
  ontolex:lexicalForm
    :fm_fudbalska_utakmica_2309936,
    :fm_fudbalske_utakmice_2309938.
    :fm_fudbalskoj_utakmici_2309942,
    ...
# inflected forms
:fm_fudbalskoj_utakmici_2309942
  a ontolex:Form;
  ontolex:writtenRep
    "fudbalskoj utakmici"@sr;
  morph:grammaticalMeaning
    [lexinfo:gender lexinfo:feminine;
    lexinfo:number lexinfo:singuar;
    lexinfo:case lexinfo:acusative;
    lexinfo:animacy lexinfo:inanimate];
  morph:grammaticalMeaning
    [lexinfo:gender lexinfo:feminine;
    lexinfo:number lexinfo:singuar;
    lexinfo:case lexinfo:locative;
    lexinfo:animacy lexinfo:inanimate].

...
```

4.2 OntoLex-FrAC Part of FudLe

The Ontolex Module for Frequency, Attestation, and Corpus Information (FrAC) is still under development and in this paper, we are relying on a Draft Community Group Report.[12] Due to the potential changes in the FrAC model, the modeling examples presented may be subject to modifications in future development.

The auxiliary class `:SrFudKo` is defined to provide convenient handling and shorter notation. Currently, the version of the corpus *srFudKo* published

[12]https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md accessed 7.8.2023

in the *noSketch engine* (Kilgarriff et al., 2014) instance is managed by the Society for Language Resources and Technologies - JeRTeH, is linked.[13]

We introduce specialized sub-classes for the two frequency types: `:SrFudKo_token_freq`, for inflected forms frequency and `:SrFudKo_lemma_freq` for a total of all inflected form-frequencies of a lexical entry. Just to mention that in this case: the "token" can be also a multi-word unit. This represents a more compact encoding, as the data does not have to be repeated for each individual observable.

```
# football corpus
:SrFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
  owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/
  #dashboard?corpname=FudKo>] .
:SrFudKo_token_freq rdfs:subClassOf
    frac:Frequency, :SrFudKo,
    [a owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "token frequency"].

# general language corpus
:SrpKor2021 a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
  owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/
  #dashboard?corpname=SrpKor2021>] .
:SrpKor2021_token_freq rdfs:subClassOf
    frac:Frequency, :SrpKor2021,
    [a owl:Restriction;
    owl:onProperty dct:description;
    owl:hasValue "token frequency"].
```

...

Let us notice that absolute and relative (per million) frequencies from several corpora are available in the lexical database for (simple) words. Figure 1 shows that the information about the frequency class: top 100, 500, 1000, etc. is available as well. It can be seen that the lemma *utakmica* is in the top 100 most frequent lemmas in the SrFudKo corpus with a relative frequency of 955.23 per million and in the top 1000 most frequent in SrpKor2021 corpus with a relative frequency of 47.13 per million. The absolute frequencies for the inflected form (token) *utakmici* and lexical entry (lemma) *utakmica* are encoded as follows:

```
# inflected form frequency
:fm_utakmici_4569855 frac:frequency
    [a :SrFudKo_token_freq;
    rdf:value "3739"].
 :fm_utakmici_4569855 frac:frequency
    [a :SrpKor2021_token_freq;
    rdf:value "23055"].

# lemma frequency
:le_utakmica_38171
    [a :SrFudKo_token_freq;
    rdf:value "29479" ] .
:le_utakmica_38171
    [a :SrpKor2021_token_freq;
    rdf:value "138573" ] .
```

In terms of multi-word units, absolute frequencies are retrieved using the CQL (Corpus Query Language) expressions, while relative frequencies are calculated by dividing the headword frequency.

The dilemma in terms of frequencies was related to the multi-word expressions frequency: whether or not the same property should be used `SrFudKo_token_freq` or it should be introduced the `SrFudKo_mwe_freq`. The possible solution may be the following:

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

Furthermore, the frequencies are given for the multi-word inflected forms *fudbalskoj utakmici* and the multi-word lexical entry *fudbalska utakmica*.

```
# mwe inflected form frequency
:fm_fudbalskoj_utakmici_2309942
  frac:frequency
    [a :SrFudKo_mwe_freq ;
    rdf:value "38" ] ;
  frac:head :fm_utakmici_4569855 .
:fm_fudbalskoj_utakmici_2309942
  frac:frequency
    [a :SrpKor2021_mwe_freq ;
    rdf:value "495" ] ;
  frac:head :fm_utakmici_4569855 .

# mwe lemma frequency
:le_fudbalska_utakmica_220902
  frac:frequency
    [a :SrFudKo_mwe_freq;
    rdf:value "219"];
  frac:head
    :le_utakmica_38171 .
:le_fudbalska_utakmica_220902
  frac:frequency
    [a :SrpKor2021_mwe_freq;
    rdf:value "2749"];
  frac:head
    :le_utakmica_38171 .
```

The attestation example *"Odavno na Banovom brdu nije bilo toliko gledalaca na jednoj fudbalskoj utakmici."*, translated to English: *"It has been a long time since there were so many spectators at one football match at Banovo Brdo"* is encoded by using properties `frac:attestation` and `frac:quotation`. It has been added manually, but automatizing the process is expected in the future:

```
# single word inflected form attestation
:fm_utakmice_4569854 [
```

[13] https://jerteh.rs/

```
   frac:quotation "Gledao sam sve
   utakmice tih timova .";
   frac:observedIn :SrFudKo].
:fm_utakmice_4569854 [
   frac:quotation "Mi u ovoj vašoj
   utakmici, u vašoj trgovini,
   nećemo da učestvujemo ..";
   frac:observedIn :SrpKor2021].

# multiword inflected form attestation
:fm_fudbalskoj_utakmici_2309942
   frac:attestation [
   frac:quotation "Odavno na Banovom
     brdu nije bilo toliko gledalaca
     na jednoj fudbalskoj utakmici.";
   frac:observedIn :SrFudKo].
:fm_fudbalskoj_utakmici_2309942
   frac:attestation [
   frac:quotation "Gospodo, ponašajte se
   pristojno, nije ovo fudbalska utakmica
   , ovo je parlament Srbije .";
   frac:observedIn :SrpKor2021].
```

5 Conclusion

The initial results of the ongoing activity in the creation of the Serbian language Football dictionary are presented, fully proving that it is possible to semi-automatically generate lists of terms and football expressions for the Serbian language. The corpus-driven approach is complemented by manual evaluation and classification of term entries. Current activities include 1) refining the produced data set through additional semantic annotation inspired by the *Kicktionary* (Schmidt, 2009) project, 2) automatic morphological inflection, which is followed by manual evaluation of the morphological classes for all new multi-word units, 3) refining the exporting procedures from the *LeXimirka* database to the *ttl*, 4) the automatic selection of good corpus examples, 5) including footballing terms' derivation and variation, and ultimately 6) word embedding integration. We will follow the initiatives related to the improvement of terminology modules for Ontolex and improve our resources according to new specifications.

Acknowledgements

This paper is partially supported by the COST Action NexusLinguarum – "European network for Web-centred linguistic data science" (CA18209), supported by COST (European Cooperation in Science and Technology).

References

Gunnar Bergh and Sölve Ohlander. 2012. Free kicks, dribblers and wags. exploring the language of "the people's game". *Moderna språk*, 106(1):11–46.

Gunnar Bergh and Sölve Ohlander. 2019. A hundred years of football english: A dictionary study on the relationship of a special language to general language. *Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses*, 32:15–43.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Unifying morphology resources with ontolex-morph. a case study in german. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational morphology with ontolex-morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86.

Christian Chiarcos, 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*, pages 1–9.

Rove Luiza De Oliveira Chishman, Aline Nardes dos Santos, Diego Spader de Souza, and João Gabriel Padilha. 2015. The relevance of the sketch engine software to build field–football expressions dictionary. *Revista de Estudos da Linguagem*, 23(3):769–796.

Serge Heiden. 2010. The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *24th Pacific Asia conference on language, information and computation*, volume 2–3, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.

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

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

Olivera Kitanović, 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.

Bettina Klimek, John P McCrae, Julia Bosque-Gil, Maxim Ionov, James K Tauber, and Christian Chiarcos. 2019. Challenges for the representation of morphology in ontology lexicons. *Proceedings of eLex*.

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

Cvetana Krstev, 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, 2015*.

Cvetana Krstev and Duško Vitas. 2009. An effective method for developing a comprehensive morphological e-dictionary of compounds. In *Proceedings of Lexis and Grammar Conference, Bergen*, pages 204–212.

Biljana Lazić and Mihailo Škorić. 2020. From dela based dictionary to leximirka lexical database. *Infotheca*.

Wojciech Liponski. 2009. "hey, ref! go, milk the canaries!" on the distinctiveness of the language of sport. *Studies in Physical Culture and Tourism*, 16(1).

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

Aleksandar Mihajlović. 2003. *Fudbalski rečnik / Footbal Dictionary / Dictionnaire du football / Diccionario del fútbol*. A. Mihajlović, Belgrade.

Gosse Minnema. 2021. Kicktionary-lome: a domain-specific multilingual frame semantic parsing model for football language. *arXiv preprint arXiv:2108.05575*.

Andjelka Pejovic. 2021. Logros lexicográficos del hispanismo serbio y el croata. *Revista de Lexicografía*, 26:113–130.

Roger Penn. 2016. Football talk: sociological reflections on the dialectics of language and football. *European Journal for Sport and Society*, 13(2):154–166.

Agata Savary, Cvetana Krstev, and Duško Vitas. 2007. Inflectional non compositionality and variation of compounds in french, polish and serbian, and their automatic processing. *Bulag-Bulletin de Linguistique Appliquée et Générale*, 32:73–94.

Thomas Schmidt. 2009. The kicktionary–a multilingual lexical resource of football language. In *Multilingual FrameNets in computational lexicography: methods and applications*, pages 101–132. de Gruyter.

Ranka Stanković, 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, Miyazaki, Japan, May 7-12, 2018*.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016*.

Ranka Stanković, Cvetana Krstev, Mihailo Škorić, Rada Stijović, and Nebojša Vasiljević. 2021. Towards automatic definition extraction for serbian. *Proceedings of the XIX EURALEX Congress of the European Association for Lexicography: Lexicography for Inclusion (Volume 2). 7-9 September (virtual)*, pages 695–704.

Ranka Stanković, 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, October 2011, Jachranka, Poland*, pages 77–84.

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

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In *Proc. of The 12th LREC*, pages 3947–3955, Marseille, France. European Language Resources Association.

Miloš Utvić. 2011. Annotating the corpus of contemporary serbian. In *Proceedings of the INFOtheca '12 Conference*, pages 36–47.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, XVIII:279–292.

Jovan Čudomirović. 2014. Mobilizacija publike: Izveštavanje dnevnih novina u srbiji o nastupima fudbalske reprezentacije. *Zbornik Matice srpske za filologiju i lingvistiku*, 57(2):143–159.