
Named Entity Recognition and Relation Extraction in Serbian Football Text

Jelena Lazarević¹[0009-0004-4481-2729], Ranka Stanković²[0000-0001-5123-6273],
Olivera Kitanović²[0000-0002-7571-2729]

¹ Language Resources and Technologies Society – JeRTeh, Serbia,
jelazarevic1@gmail.com

²University of Belgrade, Faculty of Mining and Geology, Serbia,
{ranka.stankovic|olivera.kitanovic}@rgf.bg.ac.rs

Scientific paper

DOI: 10.18485/judig.2026.2.ch7

Abstract

The language of football represents a highly dynamic and semantically rich domain for Natural Language Processing (NLP), particularly for tasks such as: Named Entity Recognition (NER), Named Entity Linking (NEL), and Relation Extraction (RE). The discourse of football contains numerous entities and complex semantic relations including: players, clubs, coaches, competitions, transfers, injuries, and match events. This paper focuses on information extraction from the Serbian football corpus *SrFudKo*, originally developed for contrastive discourse analysis and corpus-based linguistic research. Previous work resulted in the construction of Serbian and Spanish football corpora, while semantic annotation tasks have not yet been systematically explored, making this precious linguistic resource as a basis for the knowledge graph construction as well as semantic search. The goal of this research is to enrich the Serbian football corpus with semantic annotations and structured relational information. The paper investigates the methods for recognizing football-specific named entities, linking them to external knowledge bases, such as Wikidata, and extracting semantic relations between entities. Paying special attention to the football-specific relation types, we included the relations between players, clubs, coaches, competitions, and events, as well as their organization within a domain-oriented ontology. In addition to presenting the corpus and the proposed ontology, the paper provides quantitative analysis and illustrative examples of extracted entities and their relations, based on the Serbian football-themed articles. The presented approach contributes to the development of semantically enriched football resources, suitable for knowledge graph population, and domain-specific NLP applications.

Keywords: football, corpus, NER, NEL, relation extraction, Serbian

1. Introduction

The language of football represents a highly dynamic and information-rich domain for Natural Language Processing (NLP), particularly for Named Entity Recognition (NER), Named Entity Linking (NEL), and Relation Extraction (RE). Football texts contain numerous entities and semantically rich relations involving players, clubs, coaches, competitions, transfers, injuries, and tactical events. Such information appears in sports journalism, live commentary, match reports, fan discussions, social media, and statistical platforms.

This paper focuses on the extraction of football-related entities and relations from Serbian domain-specific corpus. Previous research resulted in the development of Serbian and Spanish football corpora primarily intended for contrastive discourse analysis and corpus-based linguistic research (Lazarević, 2024). These resources provide valuable multilingual foundation for studying the football discourse, and through its development and continuous research, by applying NER, NEL, and RE we further enrich this corpus, and enable the implementation of NLP tasks, knowledge graph construction, and information extraction (IE).

The aim of this paper is therefore to extend the existing Serbian football corpus *SrFudKo* (Lazarević, 2025) with semantic annotations and structured relational information. In particular, the paper defines and applies methods for identifying football-specific named entities, linking them to external knowledge bases such as Wikidata, and extracting semantic relations between entities. By doing so, the paper contributes toward the creation of richer multilingual football resources even more suitable for corpus linguistics, semantic search, knowledge graph population, and domain-specific NLP applications.

The remainder of the paper is organized as follows. Section 2. presents an overview of related research, with particular emphasis on relation extraction approaches and frame-semantic resources for the football domain. Section 3. introduces the Serbian football corpus *SrFudKo* (Lazarević, 2025), describing their structure, compilation principles, and their role as comparable multilingual resources. Section 4. is dedicated to IE, starting with Section 4.1. with NER, followed by Section 4.2. that defines the relation types considered

in this research, including relations between players, clubs, coaches, competitions, and events, together with their organization according to entity pairs, continuing in Section 4.3. with core of ontology of football-specific organization. Section 5. presents first examples of extractions in Section 5.1., quantitative information regarding extracted entities and relations using rule-based approach in Section 5.2., using LLM-based approach in Section 5.3., demonstrating the linguistic and semantic complexity of football discourse in Serbian. Research perspectives and possible extensions of the proposed ontology are discussed in concluding remarks in Section 6..

2. Related Research

2.1. Relation Extraction

RE represents one of the central tasks in IE and, more broadly, in NLP. The goal of RE is to identify semantic relations between entities mentioned in text and to transform unstructured textual content into structured relational knowledge (Zhao et al., 2024). Such structured representations are essential for numerous downstream applications, including knowledge graph construction, question answering, semantic search, recommendation systems, and information retrieval (Detroja et al., 2023). The quality and effectiveness of these applications largely depend on the availability of reliable annotated datasets used for the training and evaluation of extraction models.

Hendrickx et al. (2010) introduced the SemEval-2010 Task 8 dataset for multi-way classification of semantic relations between pairs of nominals, providing one of the most widely used benchmarks for sentence-level RE. Most recent advances in RE have focused primarily on English due to the availability of large-scale annotated datasets, but Serbian language remains significantly under-resourced when it comes to semantically annotated corpora and relation extraction resources. The manual creation of RE datasets remains expensive, time-consuming, and difficult to scale, despite manual annotation still being considered the gold standard for corpus quality (Diaz-Garcia and Lopez, 2025).

Recently, Pinto et al. (2026) introduced RelEx-PT, a balanced sentence-level Portuguese relation extraction dataset covering 18 Wikidata-derived relation types. The dataset was created through a distant supervision pipeline aligning Wikidata triples with Por-

tuguese Wikipedia sentences and refined using NLI-based filtering. The authors evaluated the dataset in relation classification, relation triple extraction, and open IE settings using both prompting and fine-tuning approaches with large language models (LLMs). Among other general relations they introduced the relation of sport, between a person or event, and the associated sport.¹

Ettaleb et al. (2026) introduced ReTaT, a collection of manually annotated datasets designed for joint relation extraction from textual paragraphs and tables extracted from Wikipedia pages. The authors constructed three domain-specific corpora covering Business, Female Celebrities, and Telecommunications, with annotations focused on semantic relations distributed across both textual and tabular structures. The relations were aligned with Wikidata properties, while entities were linked to Wikidata QIDs to ensure consistency across modalities. The authors evaluated the datasets using instruction-based prompting with large language models (LLMs), and demonstrated that joint processing of textual and tabular information substantially improves RE performance compared to single-source extraction approaches.

2.2. Frame Semantics and the Kicktionary Resource

Frame Semantics represents a linguistic approach focused on connecting word meaning with the syntactic and semantic contexts in which these words appear. According to Fillmore (1976), lexical meaning is understood through structured experiential contexts called *frames*, which organize knowledge about events, participants, and relations. Words are therefore interpreted not in isolation, but through previously acquired conceptual experiences, associated with particular situations. This approach enables the analysis of lexical meaning in natural contexts and supports the study of cognitive mechanisms underlying language understanding and its use.

Frame semantics has proven particularly influential in corpus linguistics and lexical semantics because it provides a systematic and empirically grounded framework for describing word usage. By modeling semantic frames and their participants, researchers can investigate lexical, syntactic, discourse, and sociolinguistic phenomena in authentic language data (Fillmore et al., 2003). The theoretical frame-

¹ <https://github.com/TomasCCPinto/RelEx-PT/blob/main/prompts/RC/Descriptive.md>

work proposed by Fillmore (1976) was operationalized through the *FrameNet* project (Fillmore et al., 2003), which aimed to construct a large-scale lexical database based on frame semantics. The project combined insights from lexical semantics, corpus linguistics, hyper-text technologies, and natural language processing in order to create a digital lexical resource richer and more functional than traditional dictionaries (Baker, 2012).

One of the most prominent domain-specific applications of Frame semantics is *Kicktionary*², a multilingual electronic dictionary of football language covering English, German, and French (Schmidt, 2009). The resource contains 1,926 football-related lexical units, including 599 English, 792 German, and 535 French terms. Kicktionary organizes terminology into a hierarchy of scenarios and semantic frames, each associated with authentic examples extracted from written and spoken football discourse. The resource was primarily designed to support understanding, translation, and paraphrasing of football texts, while also serving as an experimental platform for studying the application of lexical-semantic theories and corpus-based methodologies to specialized domains (Minnema, 2021). Minnema and Nissim (2021) proposed a flexible sequence-labeling architecture based on BERT that supports both independent and jointly integrated prediction of semantic frames and roles. The proposed configurations achieved performance comparable to more complex traditional approaches, while simultaneously providing greater transparency regarding the interactions between pipeline components.³

Kicktionary includes 16 scenarios and 104 semantic frames, with 8,164 annotated examples distributed across the three languages. Lexical units (LUs) represent words and expressions used to describe entities, actions, and events within a frame, such as *ball*, *goal*, *attack*, or *player*. These lexical units are grouped into 552 synsets connected through semantic relations such as hypernymy/hyponymy, meronymy, and troponymy. The resource therefore combines FrameNet-style semantic organization with lexical-semantic relations similar to those found in WordNet-like resources.

An important characteristic of Kicktionary is its hierarchical organization from scenarios to frames and finally to lexical units. For example, the scenario *Chance* includes several related frames such

² <http://www.kicktionary.de/>

³ <https://gitlab.com/gosseminnema/bert-for-framenet/>

as *Create_Chance*, *Miss_Chance*, and *Convert_Chance*. Each frame specifies its frame elements (semantic roles), such as *team*, *player*, *opportunity*, or *shot*, together with multilingual lexical realizations and annotated usage examples. This structure allows detailed modeling of football events and participant roles within authentic discourse.

Although originally conceived as a lexical resource, Kicktionary has also been employed for automatic semantic frame annotation. Minnema (2021) adapted the LOME semantic frame parsing model (Xia et al., 2021) for automatic annotation of football texts according to Kicktionary frames and semantic roles. Despite the relatively limited amount of annotated training data, the best-performing model achieved an F1-score of 0.83 for frame prediction and 0.81 for semantic role prediction. These results demonstrate the applicability of frame-semantic resources not only for lexicographic and linguistic purposes, but also for downstream NLP tasks such as semantic parsing and information extraction.

Considering the success of FrameNet-inspired resources for major languages and specialized domains, as well as the relatively limited work on frame-semantic resources for Serbian (Marković et al., 2021), the Kicktionary framework provides valuable inspiration for the development of Serbian domain-oriented semantic resources and annotation methodologies.

3. Serbian Football Corpus *SrFudKo*

The corpus *SrFudKo* is developed from the sports sections of news media sites Blic, B92, and Politika, as well as from specialized sporting news sites Sportklub and Mondo (Lazarević, 2024; Lazarević, 2025). After filtering and cleaning, the Serbian football corpus *SrFudKo* contains 10,100,553 tokens, including 8,618,426 words and 1,482,127 punctuation marks. Each token is annotated with its lemma (attribute *srlemma*) and part-of-speech tag (attribute *srpos*) using approach described in (Stanković et al., 2020).

Each corpus file is represented as an XML document composed of six structural units: `<data>` (root element), `<date>` (publication date), `<p>` (paragraph or article text), `<post>` (published article), `text` (complete document with metadata), and `title` (article title). The files are organized by publication year and by the news portal from which the articles were collected, resulting in a total of 37 XML files. Overall, the corpus contains 11,117 football-related articles.

Each article is consistently structured into three main components: title, publication date, and article body.

The distribution of word classes within *SrFudKo* reflects the characteristics of football journalism and sports discourse. Using the Universal Dependencies annotation scheme, common nouns (NOUN) constitute the largest category with 22.5%, followed by verbs (VERB) with 11.8%, adpositions (ADP) with 11%, and proper nouns (PROPN) with 9.5%. The relatively high proportion of proper nouns is expected due to the frequent occurrence of player names, clubs, competitions, stadiums, and geographical locations in football texts (Lazarević, 2025).

For this research, based on *SrFudKo*, a subset of 1,7M tokens was selected, covering period from 2008 to 2021 from Politika and Mondo portals, having 77,299 sentences with 52,624 PERS, 33,930 ORG, 23,477 LOC and 4,878 EVENT (Nešić et al., 2026). For relation selection, the inspiration was used from Kicktionary dictionary based on frame semantics (Minnema, 2021); previously thoroughly and extensively covered by Lazarević (2025).

4. Information Extraction

4.1. Named Entities in Football Text

Football texts contain several important named entity categories, including:

- PERS – persons: football players, coaches, referees, managers
- ORG – organizations: football clubs and national teams, leagues, cups, and championships, football federations, media organizations
- EVENT – specific football matches or fixtures
- LOC – stadiums, cities, and countries
- DATE and TIME – match dates and schedules

Examples from Serbian football texts include ORG entities such as: Partizan, Crvena zvezda, or PERS: Dušan Vlahović, Dragan Stojković Piksi, or LOC: Marakana.

4.2. Relation Types in Football Corpora

Football corpora contain numerous relation types that are highly suitable for information extraction and knowledge graph construction. Wikidata has become an important resource in football-

related NLP and knowledge graph research, providing structured and multilingual information about players, clubs, competitions, stadiums, transfers, and match events, which enables entity linking, semantic annotation, and relation extraction in football texts (Vrandečić and Krötzsch, 2014). In this section, we will elaborate on the relation types in football corpora and its mapping to Wikidata properties. In Wikidata, an item has identifier starting with Q, for example: Q188656 for FK Partizan - association football club. The identifiers for property starts with P, for example P54 (member of sports team).

Player and club relations are among the most frequent and semantically important relation categories. We list some of them with associated identifiers from Wikidata is mapping could be established:

- `playsFor(Player, Club) ↦ P54`,
- `transferredTo(Player, Club)`
- `loanedTo(Player, Club)`
- `captainOf(Player, Club) ↦ P634`
- `coachedBy(Player, Coach) ↦ P286`
- `teamPosition(Player, Team) ↦ 413`

For example, in the sentence: *Aleksandar Mitrović je srpski profesionalni fudbaler koji igra na poziciji napadača za klub Fulem i reprezentaciju Srbije*. (‘Aleksandar Mitrović is a Serbian professional football player who plays as a striker for Fulham and the Serbia national football team.’) several relations can be observed:

```
playsFor(Aleksandar Mitrović, Fulem)
playsFor(Aleksandar Mitrović, reprezentacija Srbije)
position(Aleksandar Mitrović, napadač)
nationality(Aleksandar Mitrović, srpski)
```

This sentence contains several explicit football-based relations. The verb phrase *igra...za* (‘plays for’) triggers the relation `playsFor` between the player and both the club and national team. The phrase *igra na poziciji napadača* (‘plays as a striker’) expresses the player’s position, while the adjective *srpski* (‘Serbian’) indicates nationality information.

Match-event relations are central for event extraction and sports analytics.

- `scoredGoal(Player, Match)`
- `assistedGoal(Player, Player)`
- `receivedCard(Player, CardType)`
- `injuredIn(Player, Match)`

- participatedIn(Team, Match)
- wonMatch(Team, Team)
- lostTo(Team, Team)

Serbian football reporting often contains highly event-oriented descriptions such as: *Crvena zvezda je pobedila Partizan u večitom derbiju*. (‘Crvena zvezda won against Partizan in the Eternal Derby.’)

```
wonMatch(Crvena zvezda, Partizan)
participatedIn(Crvena zvezda, Eternal Derby)
participatedIn(Partizan, Eternal Derby)
```

The verb *pobediti* (‘win’) expresses the outcome of the football match, indicating that *Crvena zvezda* was the winning team and *Partizan* the defeated opponent. The phrase *Eternal Derby* denotes the traditional rivalry match between the two clubs.

Cup (competition) relations are useful for long-term football knowledge bases and historical analysis.

- wonCup(Team, Cup)
- qualifiedFor(Team, Cup)
- eliminatedBy(Team, Team)
- hostedBy(Cup, Country)

The player attribute and career relations are especially useful for player profiling and scouting applications.

- bornIn(Player, Location)
- nationality(Player, Country)
- position(Player, Position)
- debutFor(Player, Club)
- retiredFrom(Player, Club)
- contractUntil(Player, Date)
- marketValue(Player, Value)

Advanced football analytics often require the extraction of tactical and analytical relations that may not be explicitly stated.

- pressedBy(Team, Team)
- markedBy(Player, Player)
- formedPartnership(Player, Player)
- playedFormation(Team, Formation)
- dominatedPossession(Team, Match)

These relations are more difficult because they often rely on implicit interpretation and contextual inference rather than direct lexical triggers.

Financial and transfer-related relations are particularly important for sports business intelligence and media analysis.

- `signedFrom(Player, Club)`
- `transferFee(Player, Amount)`
- `contractExtended(Player, Club)`
- `sponsoredBy(Club, Brand)`

Social media and fan discourse provide us with additional relation types involving sentiment and rivalry. In the Serbian football discourse, rivalry relations are especially prominent: Crvena zvezda, the rival of Partizan.

- `praised(Player, Coach)`
- `criticized(Player, Media)`
- `supportedBy(Club, FanGroup)`
- `rivalOf(Club, Club)`

4.3. Ontology Organization by Entity Pairs

A practical ontology for football relation extraction can be organized according to entity pair types, as presented in Table 1.

Entity1	Entity2	Example relations
PERS	ORG _{TEAM}	playsFor, transferredTo, captainOf
ORG _{TEAM}	PERS _{COACH}	managedBy
PERS	PERS	assistedGoal, substituted, trainedBy
ORG _{TEAM}	ORG _{TEAM}	defeated, rivalOf
PERS	EVENT _{MATCH}	scoredIn, injuredIn
ORG _{TEAM}	EVENT _{CUP}	wonCup, qualifiedFor
PERS	LOC	bornIn
ORG _{CLUB}	ORG _{BRAND}	sponsoredBy

Table 1: Examples of football relation types organized by entity pairs

Table 2 presents a set of Serbian trigger lexemes and expressions associated with exemplified of football-specific semantic relations used in the relation extraction process. The listed triggers represent typical lexical indicators that frequently occur in football discourse and may signal relations between players, clubs, coaches, competitions, events, and locations. Since football journalism relies heavily on repetitive domain-specific phraseology, expressions such as *igrati za* (play for), *postići gol* (score a goal), or *plasirati se* (qualify) provide valuable linguistic cues for identifying semantic relations in text. The

inclusion of English translations facilitates interpretation within the context of an English-language paper and enables easier comparison with similar multilingual relation extraction studies. These trigger lexemes can additionally serve as a basis for constructing rule-based patterns, prompt construction and semi-automatic extraction of football relations from Serbian texts.

The query patterns should combine previously defined entity constraints shown in Table 1 and lexical triggers for relation extraction. For example, the relation `playsFor` is expected to occur between entities of types `PERS` and `ORG_TEAM`, meaning that candidate sentences should contain both a person entity and a football club entity in addition to the corresponding lexical trigger. Such constraints significantly reduce noise and improve the precision of candidate retrieval. However, due to the linguistic variability and semantic complexity of football discourse, purely rule-based approaches often produce either overly restrictive patterns with low recall or overly general patterns with reduced precision.

For this reason, we adopted a hybrid approach. In the first step, the candidate sentences with at least two entities of interest annotated, were then selected, and in the second, the potential football relations were identified and classified using a rule-based and prompt-engineering approach based on LLMs. In this setup, NER serves as an efficient filtering mechanism that narrows the search space and retrieves semantically relevant contexts, whereas the LLM performs deeper semantic interpretation and determines the exact relation type between the detected entities. This combination enables both scalable corpus exploration and more robust semantic relation extraction in morphologically rich football texts.

5. Results

5.1. Examples of Relation Extraction

From 77,299 sentences, in first step a total of 27,442 was extracted, where at least two entities of interest were annotated, having in mind that we are interested for relation extraction between players, teams, coaches, matches, and cups, that are explicitly expressed in sports reports.

Before proceeding, preliminary testing of possible relation that can be found in text was performed, as well as elaborating different

Relation	Serbian trigger expressions
playsFor	igrati za (play for), nastupati za (perform/play for), član (member), nositi dres (wear the jersey)
transferredTo	preći u (transfer to), potpisati za (sign for), pojačati (strengthen/join), transfer (transfer), stići u (arrive at)
captainOf	kapiten (captain), nositi kapitensku traku (wear captain's armband), predvoditi tim (lead team)
managedBy	trenner (coach), voditi ekipu (manage the team), trenirati tim (coach/train the team), šef stručnog štaba (head coach)
assistedGoal	asistirati (assist), namestiti gol (set up a goal), dodati za gol (pass for a goal), upisati asistenciju (record an assist)
substituted	zameniti (substitute), ući umesto (come instead of), izaći iz igre (leave the game), zamena (substitution)
trainedBy	trenirati kod (train under), raditi sa (work with), pod vođstvom (under leadership of), pod komandom (under the command of)
defeated	pobediti (defeat), savladati (overcome/beat), nadigrati (outplay), slaviti protiv (win against)
rivalOf	rival (rival), rivalstvo (rivalry), derbi (derby), večiti rival (eternal rival)
scoredIn	postići gol (score a goal), dati gol (score), pogoditi (hit/score), zatresti mrežu (find the net)
injuredIn	povrediti se (get injured), doživeti povredu (suffer an injury), napustiti teren (leave the pitch)
wonCup	osvojiti kup (win the cup), podići trofej (lift trophy), trijumfovati (triumph), osvojiti titulu (win the title)
qualifiedFor	plasirati se (qualify), izboriti plasman (secure qualification), kvalifikovati se (qualify)
bornIn	rođen u (born in), poreklom iz (originating from), rodno mesto (birthplace)
sponsoredBy	sponzor (sponsor), generalni sponzor (main sponsor), partner (partner), tehnički partner (technical sponsor)

Table 2: Serbian triggers for football relation extraction.

annotation and formatting of extracted triplets entity-relation-entity. Below, we present several examples from the Serbian football discourse, alongside the possible relation annotations.

Example 1. *[PERS Dušan Vlahović] je postigao [EVENT gol] protiv [ORG Milana] u [TIME 78. minutu].* (*[PERS Dušan Vlahović] scored a [EVENT goal] against [ORG Milan] in the [TIME 78th minute].*)

```
scoredGoal(Dušan Vlahović, goal_event_1)
scoredAgainst(Dušan Vlahović, Milan)
eventTime(goal_event_1, 78. minut)
```

In this example, the predicate *postigao gol* explicitly signals a scoring event. The player is extracted as the scorer, while *Milan* is interpreted as the opposing team. The temporal expression *u 78. minutu* provides additional event information.

Example 2. *[ORG Crvena zvezda] je savladala [ORG Partizan] u [EVENT večitom derbiju].* (*[ORG Red Star] defeated [ORG Partizan] in the [EVENT Eternal Derby].*)

```
defeated(Crvena zvezda, Partizan)
participatedIn(Crvena zvezda, večiti_derbi)
participatedIn(Partizan, večiti_derbi)
```

The verb *savladala* indicates a match-result relation between two teams. The expression *večiti derbi* is domain-specific and refers to the rivalry match between *Crvena zvezda* and *Partizan*. This example also illustrates the importance of football-specific background knowledge.

Example 3. *[PERS Aleksandar Mitrović] je prešao iz [ORG Fulama] u [ORG Al Hilal].* (*[PERS Aleksandar Mitrović] transferred from [ORG Fulham] to [ORG Al Hilal].*)

```
transferredFrom(Aleksandar Mitrović, Fulham)
transferredTo(Aleksandar Mitrović, Al Hilal)
```

Transfer relations are typically expressed by verbs such as *preći*, *potpisati*, *pojačati*, or *napustiti* (transfer, sign, strengthen, or leave). In this sentence, the prepositional structure *iz Fulama u Al Hilal* clearly marks the source and target clubs.

Example 4. *[PERS Dragan Stojković Piksi] predvodio je [ORG reprezentaciju Srbije] na [EVENT Svetskom prvenstvu].*

([_{PERS}Dragan Stojković Piksi] led the [_{ORG}Serbian national team] at the [_{EVENT}World Cup].)

managedBy(reprezentacija Srbije, Dragan Stojković Piksi)
participatedIn(reprezentacija Srbije, Svetsko prvenstvo)

Here, the verb *predvodio* (led) expresses a coach/team relation. The example is interesting because the coach is mentioned first, while the managed team appears as the object of the verb. This shows that the relation extraction must take syntactic structure into account, not only the entity order.

Example 5. [_{ORG}Partizan] je produžio ugovor sa mladim napadačem do [_{TIME}2027. godine]. ([_{ORG}Partizan] extended the contract with the young striker until [_{TIME}2027].)

contractExtended(Partizan, mladi_napadač)
contractUntil(mladi_napadač, 2027)

This example illustrates a contract-related relation. If the player is not named explicitly, the system may extract a partial relation with an underspecified argument, such as *mladi_napadač* (young striker). Such cases are frequent in sports journalism and require coreference resolution or linking with previous sentences.

Example 6. [_{ORG}Real Madrid] je eliminisao [_{ORG}Barselonu] iz [_{ORG}Lige šampiona].]

eliminated(Real Madrid, Barselona)
eliminatedFrom(Barselona, Liga šampiona)

The relation *eliminated* is triggered by the verb *eliminisao*, while the cup is introduced by the phrase *iz Lige šampiona*.

Table 3 presents several illustrative examples of lexical and syntactic patterns that may indicate football-specific semantic relations in Serbian texts. The examples demonstrate how particular trigger expressions, such as *postigao gol* (scored a goal), *nastupa za* (plays for), or *prešao iz/u* (transferred from/to), can be associated with structured semantic relations between entities. The presented patterns are intended only as representative examples rather than a complete inventory of possible football relation constructions. Due to the linguistic variability of sports journalism and the intricacy of the Serbian morphology, many additional lexical realizations and syntactic variants are possible. Consequently, these examples primarily

illustrate the general principles used for constructing rule-based extraction patterns and generating candidate sentences later processed by LLM-based relation extraction methods.

Sentence pattern	Trigger	Extracted relations
X je postigao gol protiv Y	postigao gol	scoredAgainst(X,Y)
X je asistirao X-u za gol	asistirao	assistedGoal(X,Y)
X je savladao Y	savladao	defeated(X,Y)
X je eliminisao Y iz Z	eliminisao	eliminated(X,Y), eliminatedFrom(Y,Z)
X je prešao iz Y u Z	prešao iz/u	transferredFrom(X,Y), transferredTo(X,Z)
X nastupa za Y	nastupa za	playsFor(X,Y)
X je kapiten Y	kapiten	captainOf(X,Y)
X je predvodio Y	predvodio	managedBy(Y,X)
X je trenirao Y	trenirao	trainedBy(Y,X)
X je produžio ugovor sa Y	produžio ugovor	contractExtended(X,Y)
X se plasirao u Z	plasirao se	qualifiedFor(X,Z)
X je osvojio Z	osvojio	wonCup(X,Z)
X se povredio na Z	povredio se	injuredIn(X,Z)
X je rođen u Y	rođen u	bornIn(X,Y)
Y sponzorise X	sponzorise	sponsoredBy(X,Y)

Table 3: Examples of Serbian football relation patterns and lexical triggers

5.2. Statistics for Rule-based Extraction

For the rule-based approach, we selected a subset of all relations. The execution was very fast and Table 4 reveals several important characteristics of football discourse in the Serbian corpus. The most frequent relation types are `playsFor` (3294 instances) and `managedBy` (3150 instances), indicating that football texts are strongly centered around players, clubs, and coaches, which is expected in sports journalism, with articles frequently discussing player affiliations, transfers, team lineups, coaching changes, and tactical organization.

A particularly high number of instances is also observed for the relations `substituted` (2302), `rivalOf` (2084), and `wonCup` (2022). The prominence of `substituted` reflects the event-oriented nature of match reports, where substitutions are systematically described. Similarly, `rivalOf` captures the importance of derby matches and historical rivalries between clubs, which represent a central narra-

tive component of football media discourse. Relations connected to competitions and achievements, such as `wonCup`, are also highly represented because sports journalism frequently emphasizes trophies, competition outcomes, and qualification success.

Medium-frequency relations include `captainOf`, `defeated`, and `transferredTo`. These relations correspond to recurring but more context-specific football situations, including captaincy roles, match results, and transfer activities. On the other hand, relations such as `scoredIn`, `bornIn`, and especially `qualifiedFor` appear considerably less frequently. This lower frequency may be explained by stricter lexical constraints, fewer explicit mentions in the corpus, or the limited availability of event-oriented annotations.

The statistics additionally reveal a strong dominance of relations involving `PERS` and `ORG` entities, confirming that football discourse is primarily structured around interactions between players, coaches, clubs, and teams. Relations involving `EVENT` and `LOC` entities occur less frequently, but they provide important contextual information related to competitions, matches, and geographical aspects of football competitions. Overall, the obtained distribution demonstrates that the corpus contains a rich variety of football-specific semantic structures suitable for both rule-based and LLM-based relation extraction experiments.

Relation	Entity1	Entity2	Count
<code>managedBy</code>	<code>ORG</code>	<code>PERS</code>	3150
<code>playsFor</code>	<code>PERS</code>	<code>ORG</code>	3294
<code>wonCup</code>	<code>ORG</code>	<code>EVENT</code>	2022
<code>rivalOf</code>	<code>ORG</code>	<code>ORG</code>	2084
<code>substituted</code>	<code>PERS</code>	<code>PERS</code>	2302
<code>captainOf</code>	<code>PERS</code>	<code>ORG</code>	1149
<code>defeated</code>	<code>ORG</code>	<code>ORG</code>	1088
<code>transferredTo</code>	<code>PERS</code>	<code>ORG</code>	994
<code>assistedGoal</code>	<code>PERS</code>	<code>PERS</code>	670
<code>sponsoredBy</code>	<code>ORG</code>	<code>ORG</code>	236
<code>scoredIn</code>	<code>PERS</code>	<code>EVENT</code>	137
<code>bornIn</code>	<code>PERS</code>	<code>LOC</code>	47
<code>qualifiedFor</code>	<code>ORG</code>	<code>EVENT</code>	8

Table 4: Distribution of extracted football relation candidates by entity types.

Table 5 shows notable differences in the distribution of extracted football relations across the analyzed portals. Although the *Politika* subcorpus contains a larger number of tokens and sentences, the *Mondo* portal exhibits a substantially higher density of extracted relations, both per sentence and per 1,000 tokens.

Portal	Tokens	Sentences in $() \geq 2$ ent.	Extrac. relats	Rels. per sentence	Rels. per 1K tokens
politika	965,238	43,877(13,260)	6,803	0.155	7.05
mondo	765,391	32,945(14,182)	10,378	0.315	13.56
total	1,730,629	76,822	17,181	0.224	9.93

Table 5: Statistics of extracted football relations across portals.

This difference may be explained by stylistic and genre-specific characteristics of the portals. *Mondo* articles are typically shorter, more event-oriented, and strongly focused on match reports, transfers, and football-related interactions, resulting in a higher concentration of semantic relations. In contrast, *Politika* contains more descriptive and analytical journalistic texts, which produce a lower ratio of explicit football relations. The overall statistics confirm that football discourse is highly suitable for relation extraction, especially in sports-oriented online media characterized by dense event narration and frequent references to players, clubs, competitions, and match outcomes.

5.3. Statistics for LLM-based Extraction

The Mondo subcorpus contains 6,325 sentences, of which 2,619 include two or more entities relevant for relation extraction. Due to computational costs and processing time constraints, only a subset of the corpus was used for the LLM-based extraction experiments. Processing the complete dataset would require more than 10 hours, so the present experiment was limited to the 2013 Mondo subcorpus and employed the `gpt-5.4-lite` model. The total processing time for this subset was approximately 53 minutes.

Figure 1 compares the number of extracted football relations obtained using the LLM-based and rule-based approaches. The results reveal substantial differences between the two methodologies depending on the relation type. The LLM-based approach achieved notably higher extraction counts for relations such as `playsFor`,

defeated, and especially `scoredIn`, indicating its stronger ability to capture semantically diverse and context-dependent football events. In contrast, the rule-based approach produced higher numbers for relations such as `managedBy`, `captainOf`, and `rivalOf`, where lexical triggers and syntactic constructions are more stable and predictable.

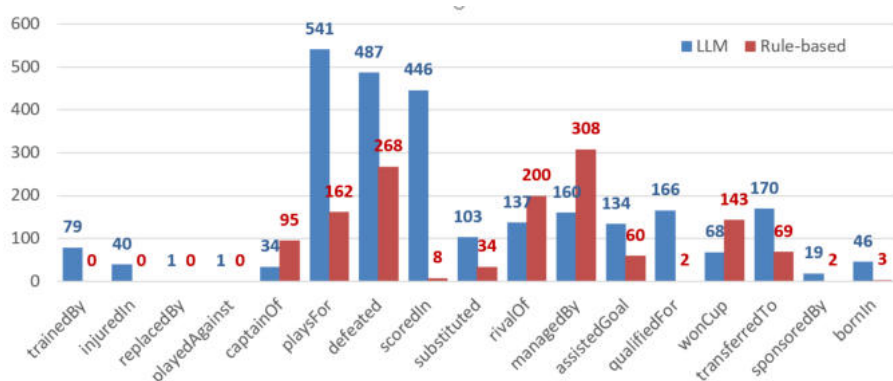


Figure 1: Comparison of relations extracted using LLM-based and rule-based approaches across different football relation types.

5.4. Evaluations and Discussion

A subset of all extractions is evaluated and of the 5,311 relation candidates, 859 (16.2%) were confirmed, 4,445 (83.7%) were rejected, and only seven remained uncertain (Table 6). The confirmation rate varied considerably across relations: `bornIn` and `qualifiedFor` had the highest rates (44.7% and 50.0%, respectively), although both had relatively few candidates. In contrast, `substitutedBy` produced the largest number of candidates (1,702), but only 5.2% were confirmed, suggesting that its extraction patterns require further refinement.

Longer sentences with many named entities generated more false positives, since system retrieved more candidates. For example in Example 7:

Example 7. *[PERSAndrija Mladenović], predstavnik za štampu vodeće stranke u [ORGVladi Srbije], ovih dana je, recimo, kao [ROLEklupski funkcioner] izneo kritiku na račun [ROLETrenera] [PERSDuška Vujoševića], izjavivši da, iako trofejni strateg ima pravo na mišljenje ne bi trebalo da se bavi optužbama na račun čelnika [ORGCrvene zvezde], nego trenerskim poslom u kojem je sjajan...* (*[PERSAndrija Mladenović]*, the spokesperson for the leading party in the *[ORGGovernment*

Table 6: Preliminary evaluation of rule-based candidates.

relation_candidate	Yes	No	Unsetain	Total
assistedGoal	45	625		670
bornIn	21	26		47
captainOf	59	262		321
defeated	81	330	2	413
managedBy	61	89	4	154
playsFor	78	141	1	220
qualifiedFor	4	4		8
rivalOf	203	355		558
scoredIn	51	86		137
sponsoredBy	11	225		236
substitutedBy	88	1614		1702
transferredTo	139	572		711
wonCup	18	116		134
Total	859	4445	7	5311

of Serbia], recently, for example, in his capacity as a [ROLEclub official], criticized [ROLEcoach] [PERSDuško Vujošević], stating that, although the trophy-winning coach is entitled to his opinion, he should focus on coaching, at which he excels, rather than making accusations against the leadership of [ORGRed Star]...)

Rule based system extracted 4 candidates from Example 7., but only the last one is correct.

```
managedBy(Vladi Srbije, Andrija Mladenović)
managedBy(Vladi Srbije, Duška Vujoševića)
managedBy(Crvene zvezde, Andrija Mladenović)
managedBy(Crvene zvezde, Duška Vujoševića) [correct]
```

An example of true positive with no false extractions is Example 8. In addition to managedBy relation, named entites are also linked to Wikidata entries.

Example 8. *Tako [PERSAleks Ferguson], jedan od najuspešnijih [PERStrenera] u istoriji svetskog fudbala, smatra da mediji mrze njegov [ORGMančester junajted].* (Thus, [PERSAlex Ferguson], one of the most successful [PERScoaches] in the history of world football, believes that the media hate his [ORGManchester United].)

```
managedBy(Mančester junajted, Aleks Ferguson)
wikidata(Mančester junajted, Q18656)
```

wikidata(Aleks Ferguson, Q44980)

A total of 354 distinct relation triples were extracted for which the corresponding relation was verified in Wikidata. Their distribution by relation type was as follows: `bornIn` (2), `captainOf` (13), `defeated` (70), `managedBy` (11), `playsFor` (34), `rivalOf` (152), `sponsoredBy` (10), `substituted` (2), `transferredTo` (2), and `wonCup` (58).

The comparison additionally demonstrates that LLMs are more effective in recognizing implicit and linguistically variable relations, while rule-based patterns provide higher precision for structurally regular constructions. Overall, the results suggest that the two approaches are complementary rather than mutually exclusive, motivating the development of hybrid pipelines combining high-recall LLM extraction with linguistically informed rule-based filtering and validation.

The evaluation of generated sentences and inclusion of more relations is ongoing activities that will be reported in final version of paper.

6. Conclusion

The football domain represents a particularly suitable environment for relation extraction, since many semantic relations between players, clubs, coaches, matches, competitions, and other football entities are explicitly expressed in sports journalism and match reports. Football discourse is characterized by repetitive domain-specific terminology, rich event structures, and clearly identifiable interactions between entities, making it highly appropriate for both rule-based and LLM-based information extraction approaches. In this paper, we presented examples of football-specific relations extracted from Serbian football discourse within the *SrFudKo* corpus, alongside the possible semantic annotations, lexical triggers, and candidate extraction patterns.

The obtained results demonstrate that football corpora provide valuable resources for the development of domain-specific NLP applications, including Named Entity Recognition, Named Entity Linking, Relation Extraction, semantic search, and knowledge graph construction. The proposed approach combines corpus-based querying techniques with prompt-based LLM extraction, enabling efficient retrieval of candidate relations while preserving semantic flexibility

necessary for handling morphologically rich Serbian texts and diverse journalistic styles.

Future research will extend the analysis to the Spanish football corpus *EsFudKo* (Lazarević, 2025), enabling comparative and cross-lingual studies between the Serbian and Spanish football discourse.

Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia under Project No. 7276, “Text Embeddings – Serbian Language Applications (TESLA)”.

References

- Collin F Baker. 2012. [FrameNet, Current Collaborations and Future Goals](#). *Language Resources and Evaluation*, 46(2):269–286.
- Kartik Detroja, CK Bhensdadia, and Brijesh S Bhatt. 2023. [A Survey on Relation Extraction](#). *Intelligent Systems with Applications*, 19:200244.
- Jose A Diaz-Garcia and Julio Amador Diaz Lopez. 2025. [A Survey on Cutting-edge Relation Extraction Techniques based on Language Models](#). *Artificial Intelligence Review*, 58(9):287.
- Mohamed Ettaleb, Thibault Ehrhart, Nathalie Aussenac-Gilles, Yoan Chabot, Mouna Kamel, Véronique Moriceau, Raphael Troncy, and Fanfu Wei. 2026. [ReTaT: A Unified Benchmark for Relation Extraction across Text and Table](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1341–1351, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Charles J Fillmore. 1976. [Frame Semantics and the Nature of Language](#). *Annals of the New York Academy of Sciences*, 280(1):20–32.
- Charles J Fillmore, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. [FrameNet in Action: The Case of Attaching](#). *International journal of lexicography*, 16(3):297–332.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. [Semeval-2010 Task 8: Multi-](#)

- way Classification of Semantic Relations Between Pairs of Nominals. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 33–38.
- Jelena Lazarević. 2025. Jezičke odlike diskursa novih medija o fudbalu: kontrastivne analiza na korpusu srpskog i španskog jezika. *University of Belgrade, Faculty of Philology*. Ph. D thesis.
- Jelena Lazarević. 2024. Tekstometrija u korpusnoj lingvistici, srpski fudbalski korpus srfudko. *Nasleđe*, (59):111–124.
- Aleksandra Marković, Ranka Stanković, Natalija Tomić, and Olivera Kitanović. 2021. FrameNet Lexical Database: Presenting a Few Frames Within the Risk Domain. *Infotheca–Journal for Digital Humanities*, 21(1):7–33.
- Gosse Minnema. 2021. Kicktionary-LOME: a Domain-specific Multilingual Frame Semantic Parsing Model for Football Language. *arXiv preprint arXiv:2108.05575*.
- Gosse Minnema and Malvina Nissim. 2021. Breeding Fillmore’s Chickens and Hatching the Eggs: Recombining Frames and Roles in Frame-semantic Parsing. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 155–165.
- Milica Ikonić Nešić, Saša Petalinkar, Olivera Kitanović, Ranka Stanković, and Miloš Utvić. 2026. CNN-based Named Entity Linking: Serbian Use Case. *Poznan Studies in Contemporary Linguistics*.
- Tomás Pinto, Catarina Silva, and Hugo Goncalo Oliveira. 2026. RelEx-PT: A Portuguese Sentence-Level Relation Extraction Dataset. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7681–7691, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Thomas Schmidt. 2009. The Kicktionary–A Multilingual Lexical Resource of Football Language. *Multilingual FrameNets in computational lexicography. Methods and applications*, pages 101–134.
- 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 *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962.

- Denny Vrandečić and Markus Krötzsch. 2014. [Wikidata: a Free Collaborative Knowledgebase](#). *Communications of the ACM*, 57(10):78–85.
- Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. [LOME: Large Ontology Multilingual Extraction](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159.
- Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. [A Comprehensive Survey on Relation Extraction: Recent Advances and new Frontiers](#). *ACM Computing Surveys*, 56(11):1–39.

Препознавање именованих ентитета и екстракција релација у српским фудбалским текстовима

Јелена Лазаревић, Ранка Станковић, Оливера Китановић

Сажетак

Језик фудбала представља веома динамичну и семантички богату област за обраду природног језика, посебно за задатке као што су препознавање именованих ентитета (Named Entity Recognition – NER), повезивање именованих ентитета (Named Entity Linking – NEL) и екстракција релација (Relation Extraction – RE). Фудбалски дискурс садржи бројне ентитете и сложене семантичке односе који укључују играче, клубове, тренере, такмичења, трансфере, повреде и догађаје са утакмица. Овај рад је усмерен на екстракцију информација из српског фудбалског корпуса *SrFudKo*, који је првобитно развијен за контрастивну анализу дискурса и корпусна лингвистичка истраживања. Претходна истраживања су резултовала формирањем српског и шпанског фудбалског корпуса, а у наставку истраживања постављени су задаци семантичког аотирања, те *SrFudKo* драгоцени лингвистички ресурс чини основу за изградњу графова знања и семантичку претрагу. Циљ овог истраживања јесте обогашивање српског фудбалског корпуса семантичким аотацијама и структурираним релационим информацијама. У раду се истражују методе за препознавање фудбалски специфичних именованих ентитета, њихово повезивање са спољним базама знања као што је Википодаци, као и екстракција семантичких релација између ентитета. Посебна пажња посвећена је типовима релација карактеристичним за фудбалски домен, укључујући односе између играча, клубова, тренера, такмичења и догађаја, као и њиховој организацији у оквиру доменски оријентисане онтологије. Поред представљања корпуса и предложене онтологије, рад доноси квантитативне лингвистичке анализе и илустративне примере екстрахованих ентитета и релација из српских фудбалских текстова. Предложени приступ доприноси развоју семантички обогашених фудбалских ресурса погодних за корпусну лингвистику, попуњавање графова знања, екстракцију информација и доменски специфичне NLP примене.

Кључне речи: фудбал, корпус, NER, NEL, екстракција релација, српски језик

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

