Contrastive Analysis of Syntax Patterns in Comparable Football Corpora in Spanish and Serbian Languages

Scientific paper DOI: 10.18485/judig.2025.1.ch10

Jelena Lazarević[1], *0009-0004-4481-2729*
Olivera Kitanović[2] *0000-0002-7571-2729*

Abstract

This paper explores the notion of Collocability as the manner in which lexical units combine with different word categories to form larger units through analyzing the semantic and syntactic principles of Serbian and Spanish football terms using comparable corpora: *srFudKo* and *esFudKo*, developed within Jelena Lazarevic's doctoral dissertation: *Language characteristics of the new media discourse on football: a contrastive analysis of the Serbian and Spanish language corpora*.

The corpus *srFudKo* includes 10,100,553 tokens from five Serbian news sites, while *esFudKo is* comprised of 9,106,812 tokens from two Spanish football websites. Both corpora are analyzed using corpus linguistics methods using CQL (Corpus Query Language) and textometric analysis. The study examines collocations based on frequency and semantic attractiveness, identifying different types, such as: adjective + noun, noun + noun, verb + noun, adverb + adjective, verb + prepositional phrase, and verb + adverb. The process of collocation extraction relies on syntactic patterns and frequency analysis. Beyond frequency, the research considers semantic closeness and context, assessing whether collocations carry specific meanings or whether they are understandable to the general public. When broadly understood, collocations transcend the footballing domain. This study also examines connections between collocations and multi-part

[1] Univerzitet u Beogradu, Filološki fakultet, doctorand, jelazarevic1@gmail.com
[2] Univerzitet u Beogradu, Rudarsko-geološki fakultet, olivera.kitanovic@rgf.bg.ac.rs

terms, showing how football-specific terms contribute to a broader understanding of the sports linguistic structure in Serbian and Spanish.

Keywords: football, corpora, terminology, collocations, Serbian, Spanish

1. Introduction

This paper aims to explore collocability as a manner in which lexical units are combined with terms from different categories, forming larger units. The research of the semantic and syntactic principles of these combinations of the Spanish and Serbian footballing terms was carried out on the comparable football corpora: *srFudKo* and *esFudko*, which were developed as part of Jelena Lazarevic's doctoral dissertation titled: *Language characteristics of the new media discourse on football: a contrastive analysis of the Serbian and Spanish language corpora.*

The football corpus of the Serbian language, named *srFudKo* was developed through texts on football from five Serbian news sites: B92, Blic, Mondo, Politika, and Sport klub, containing 10,100,553 tokens, with a total of 8,618,426 words. The corpus of Spanish-language articles on football *esFudKo*, comes from two Spanish news sites: Marca and Mundo deportivo, containing 9,106,812 tokens, with a sum of 8,024,164 words. Both corpora are located on the platform https://noske.jerteh.rs (Kilgarriff et al. 2014; Lazarević 2024) and are available to authorized users.

In this paper, the mutual lexical-semantic attractiveness of collocations is determined based on frequencies and other measures within the corpora, so that collocations are viewed in the broadest sense of Corpus linguistics - as a series of words or concepts that appear together more often than expected by mere chance. We present seven main types of collocations through the following examples: adjective + noun (fast counter), noun + noun (penalty shootout), verb + noun (to score a goal), adverb + adjective (very talented), verbs + prepositional phrase (play at the stadium) and verb + adverb (to kick hard). The process of collocation extraction represents a technique in Computational linguistics that identifies collocations within a text, or corpus of texts, using elements similar to data mining, while relying on syntactic patterns and frequencies of occurrence.

In addition to frequencies of occurrence (Levin 2008), we also consider other factors, such as semantic closeness and context in both languages. For example, do certain collocations have specific meanings, or are they only used in certain situations? We also consider whether or not the previously identified collocations are understandable to the general public who don't follow sport, meaning they aren't familiar with the language of football. In the case of a speaker from the general public, understanding them, the collocations have surpassed their origin in the football domain. They have become part of the public domain.

The contribution of this research is also reflected in analyzing the connections between collocations and multi-part terms. It is noted that the connection is strong when multi-part terms contain collocates that possess a clear meaning within the footballing domain. This helps understand the terminological connection within the language of football, providing insight into typical word combinations and their use, illustrating those that often appear in football corpora of the Serbian and Spanish languages of football. Lazarević et al. (2023) reported on football terminology, starting with the approach for compilation, followed by transformation into the OntoLex-Lemon resource.

Section 2 is dedicated to introducing the problem of syntax patterns, methods for the extraction, and an approach to contrastive analysis. Section 3 presents an overview of *srFudKo* and *esFudko*. Furthermore, Section 4 follows the contrastive analysis of illustrative and productive syntax patterns in the Spanish and Serbian languages, through collocation analysis of the most frequent nouns, the adjective + noun, and other patterns with nouns. We conclude this section with a discussion beyond frequency, about the context and semantic nuances in football collocations.

2. Methods

In Phraseology, a collocation is a type of compositional phrase, i.e. that its meaning can be understood from the words that constitute it, unlike idiomatic expressions, also known as *fixed phrases*, where the meaning of the entire unit cannot be inferred from its parts, which may be completely unrelated. There are seven main types of collocations: adjective + noun, noun + noun, noun + verb, verb + noun, adverb + adjective, verb + prepositional phrase (phrasal verbs), and verb + adverb (Prćić, 2016).

Pejović (2015) discusses collocability as the characteristic of words to combine with other words from different categories, thus forming larger units that range from phrases to sentences. Words combine according to specific semantic and syntactic principles. In Linguistic analysis, syntactic patterns represent the fundamental, recurring structures within sentences that define the way in which words are arranged and combined, according to grammatical rules. In other words, templates or patterns exist, which then determine the order and interrelation of words. For example, in the Serbian language, the pattern: adjective + noun (as in the phrase *precizan šut*/precise shot indicates that a descriptive word (adjective) naturally connects with a noun. Similarly, in Serbian, we can identify patterns such as noun + verb or verb + noun, where the syntactic structure provides the foundation for expected word usage. The terminology acquisition can be performed semi-automatically, based on lexical resources and local grammars developed for the Serbian language (Kitanović et al. 2015).

A collocation sense compositionality differentiates it from traditional idiomatic expressions, also known as fixed phrases, where the meaning of the entire unit cannot be deduced solely from the meanings of its composing parts and is often unrelated to their literal meaning. The Serbian examples from the *srFuDKo* for the seven main types of collocations are:

1. Adjective + noun: *snažan uticaj/*strong influence
2. Noun + noun: *gol linija/*goal line
3. Noun + verb: *vreme leti/*time flies
4. Verb + noun: *čuvati mrežu/*to guard the net
5. Adverb + adjective: *izuzetno dobar/*extremely good
6. Verb + prepositional phrase (phrasal verbs): *osloniti se na/*to rely on
7. Verb + adverb: *trčati brzo/*to run fast

These collocations are not arbitrary; they appear within the Serbian language as established and recognizable combinations that speakers intuitively perceive as natural and correct. Thus, syntactic patterns and collocations overlap. In their essence, collocations represent specific instances of syntactic patterns that carry additional semantic weight due to their conventionalized use.

A Concordance is a systematic representation of word and phrase use in real texts, typically within corpora, as searchable text databases. A concordance provides an overview of all occurrences of a given word or phrase, displaying its context within a sentence or a passage. Concordances enable the analysis of frequency and variations of syntactic patterns, identification of collocations through the examination of contexts where specific word combinations appear, and the evaluation of established patterns (e.g., collocations) align with expectations based on syntactic rules.

What is proven above is that syntactic patterns serve as the fundamental structural templates of a language, forming the basis on which phrases and sentences are constructed. Collocations are specific word combinations within these patterns that have become established in the language. The term Collocate is used to refer to one of the words that form part of a collocation. Thus, if *postići gol/*score a goal is a collocation, the words *postići/*score and gol/goal are collocates, meaning each word is a collocate of the other within that specific phrase.

Concordances provide an empirical insight into the use of syntactic patterns and collocations by analyzing language corpora *srFuDKo* and *esFudKo*. Through concordance analysis, we can confirm the frequency and contextual use of these patterns, ensuring that the identified collocations are not just theoretically valid but also practically present in real language use.

Following Pejović (2007, 2015), we consider collocations in their broadest sense. In this study, we define collocations as word combinations between which there is mutual lexical-semantic attraction, detected based on frequency and other corpus-based measures, without categorizing them into subtypes. Collocation extraction is a computational linguistics technique used to identify collocations in a text or corpus, relying on syntactic patterns and frequency analysis, similar to data mining techniques. Although this study doesn't focus on the typology of collocations, it is important to highlight the close relationship between collocations and multi-word terms. This connection is particularly relevant because multi-word terms often contain word combinations that qualify as collocations. Multi-word terms are typically composed of several words that regularly appear together and carry specific meaning within a particular domain (Prćić, 2016), in this case, the domain of football. This connection between collocations and multi-word terms is valuable for understanding the terminological relationships within the language of football. Through analyzing collocations frequently occurring near multi-word terms, we gain a better understanding of how these terms are used and their typical word combinations.

3. Football Corpora srFudKo and esFudKo

In the field of Corpus linguistics, the construction of specialized corpora allows for in-depth linguistic analysis of domain-specific language. This study utilizes two football-related corpora: *srFudKo*, a Serbian football corpus (Lazarević 2024), and *esFudKo*, a Spanish football corpus. These corpora provide valuable resources for examining linguistic patterns, terminology, and discourse characteristics in football news articles across the two different languages.

The *srFudKo* corpus was created using articles from five Serbian websites: B92, Blic, Mondo, Politika, and Sport Klub, tagged with Serbian taggers (Stanković et al. 2020). It is comprised of 10,100,553 tokens, of which 8,618,426 are individual words.

Figure 1 presents its partitions by web portal on the left, and by year on the right. The corpus *srFudKo* captures the stylistic and lexical features of football reporting in Serbian, offering insights into frequently used collocations, terminology, and syntactic structures prevalent in sports journalism. Given its large size and diverse sources, *srFudKo* represents a reliable dataset for analyzing football-related language use in Serbian media.

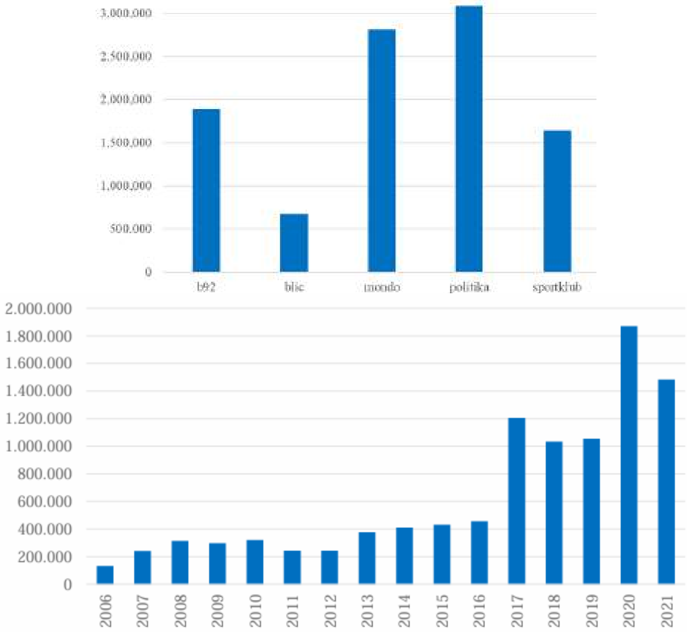


Figure 1. The partitions of Serbian football corpus *srFudKo* by web portals and years

Similarly, *esFudKo* corpus focuses on the football discourse in the Spanish language. It is compiled from articles published on the two major Spanish sports web-sites: Marca and Mundo Deportivo, tagged with TreeTagger Spanish tagset (https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt). The Spanish corpus contains 9,106,812 tokens, with 8,024,164 being words. The corpus *esFudKo* provides a comparable linguistic resource for analyzing Spanish football terminology, syntactic structures, and collocations (Lazarević & Kitanović 2024). By examining patterns within this dataset, we identify common linguistic features that define football reporting in Spanish-speaking media.

Together, these corpora serve as essential tools for comparative linguistic research. They enable cross-linguistic analysis of football discourse, highlighting similarities and differences in the ways football is discussed, reported, and conceptualized in Serbian and Spanish. Additionally, these corpora facilitate automatic extraction of collocations, frequency analysis, and syntactic pattern recognition, contributing to a broader understanding of football-specific language in different linguistic and cultural contexts.

Both corpora used in this study have undergone data extraction methods and are available on two platforms: web-based access through Noske (https://noske.jerteh.rs/) (Kilgarriff et al. 2014; Stanković et al. 2021), an online corpus analysis platform; and local access via TXM, a specialized corpus linguistics software for textual analysis (Heiden 2010;). Access to these corpora is restricted to authorized users, ensuring controlled use and data protection.

In this study, collocations are identified based on their mutual lexical-semantic attraction, determined through frequency measures and other statistical metrics applied within the corpora (Kitanović et al. 2021). This approach allows a systematic and empirical identification of word combinations that frequently occur together, ensuring that the extracted collocations are linguistically significant and relevant to the footballing language in both Serbian and Spanish.

4. Results

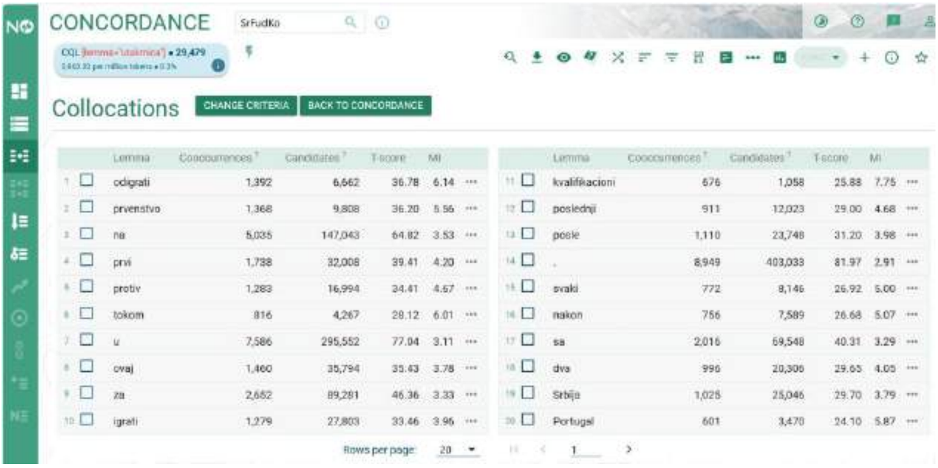
4.1. Analysis of Collocations of the Most Frequent Nouns in srFuDKo and esFuDKo

We now illustrate the analysis using the example: *utakmica/*match, a highly frequent noun in football discourse, analyzing its collocations helps us reveal typical word patterns and semantic relationships in both Serbian and Spanish football terminology. Collocations involving *utakmica* can be categorized into different structural patterns. For the pattern adjective + noun, illustrative examples are: *važna utakmica/*important match, *teška utakmica/*tough match, *prijateljska utakmica/*friendly match.

The pattern noun + noun can be illustrated with: *gol razlika/*goal difference, *revanš utakmica*/rematch. The pattern verb + noun can be illustrated with: *igrati utakmicu/*to play a match, *dobiti utakmicu/*to win a match, and *izgubiti utakmicu/*to lose a match. For the pattern noun + prepositional phrase, the illustrative examples are: *utakmica na stadionu/*match at the stadium, *utakmica pod reflektorima/*match under the floodlights, *utakmica za titulu/*match for the title.

The noun *utakmica/*match often appears in contexts describing competition, results, and game conditions. The collocations provide insight into how matches are characterized, whether by difficulty *(teška utakmica/*tough match), significance *(važna utakmica/*an important match), or type *(prijateljska utakmica/*friendly match). Additionally, the verb-noun combinations reflect actions within the game: winning (*dobiti utakmicu/*win the match*),* losing (*izgubiti utakmicu/*lose the match), or playing *(igrati utakmicu/*play the match). These structures are crucial for Computational linguistics applications such as machine translation, keyword extraction, and sentiment analysis in sports journalism.

Figure 2 presents the words (lemma) with strong connection calculated from articles as a list of collocates with the calculated statistical measures in *NoSketch engine* (Kilgarriff et al. 2014), maintained by the Language Resources and Technologies Society *JeRTeh*.



		Lemma	Cooccurrences	Candidates	T-score	MI	
1	□	odigrati	1,392	6,662	36.78	6.14	···
2	□	prvenstvo	1,368	9,808	36.20	5.56	···
3	□	na	5,035	147,043	64.82	3.53	···
4	□	prvi	1,738	32,008	39.41	4.20	···
5	□	protiv	1,283	16,994	34.41	4.67	···
6	□	tokom	816	4,267	28.12	6.01	···
7	□	u	7,586	295,552	77.04	3.11	···
8	□	ovaj	1,460	35,794	35.43	3.78	···
9	□	za	2,652	89,281	46.36	3.33	···
10	□	igrati	1,279	27,803	33.46	3.96	···
11	□	kvalifikacioni	676	1,058	25.88	7.75	···
12	□	poslednji	911	12,023	29.00	4.68	···
13	□	posle	1,110	23,748	31.20	3.98	···
14	□	.	8,949	403,033	81.97	2.91	···
15	□	svaki	772	8,146	26.92	5.00	···
16	□	nakon	756	7,589	26.68	5.07	···
17	□	sa	2,016	69,548	40.31	3.29	···
18	□	dva	996	20,306	29.65	4.05	···
19	□	Srbija	1,028	25,046	29.70	3.79	···
20	□	Portugal	601	3,470	24.10	5.87	···

Figure 2. The Collocations panel in *srFudKo* on NoSketch engine for the word *utakmica/*match

In the Spanish language, similar collocations appear with the term *partido/*match. For adjective + noun pattern: *partido importante/*important match, *partido difícil/*tough match, *partido amistoso/*friendly match. For noun + noun pattern, we what can be seen is: *final del partido/*final match, *revancha del partido/*match rematch, while for Verb + Noun: *jugar un partido/*to play a match, *ganar un partido/*to win a match, and *perder un partido/*to lose a match. Despite language differences, the collocational patterns remain structurally similar, showing the universal nature of football terminology across languages.

4.2. The Adjective + Noun Pattern in srFuDKo and esFuDKo

The adjective + noun pattern is frequently used to describe properties or characteristics of an entity. This pattern, recognized as one of the most productive in a language, was analyzed in the *srFudKo* corpus by extracting occurrences using the CQL pattern: [srpos="ADJ"] [srpos="NOUN"]. The extraction process yielded 528 candidates for combinations with a frequency higher than 100, while 84 candidates had a frequency exceeding 500. In Table 1, the frequent adjective + noun collocations related to the footballing language are listed, illustrating typical word combinations within the language of football in Serbian.

Table 1. The frequent collocations for pattern adjective + noun in the corpus srFudKo

Bigram	Freq
Crvena zvezda	11050
svetsko prvenstvo	4165
drugo poluvreme	2810
prvo poluvreme	2686
kazneni prostor	2317
fudbalski savez	2254
Evropsko prvenstvo	1808
srpski fudbal	1807
stručni štab	1708
žuti karton	1707
nacionalni tim	1647
prvi meč	1599
slobodni udarac	1478
prvi deo	1366
direktan plasman	1231
grupna faza	1214
istorijski trijumf	1121
prelazni rok	1090
mladi igrač	1051
bela tačka	1004
prvo mesto	990
crveni karton	963
dobar rezultat	939
drugi deo	834
fudbalski klub	823
veliki klub	806
velika šansa	795
dobar strelac	759
sportski direktor	758
dobra prilika	729
vezni red	727

The most frequent nouns appearing in adjective + noun structure are: *utakmica*/match (67), *igrač*/player (66), *klub*/club (62), *tim*/team (54), *gol*/goal (50), *meč*/ match|game (49), *fudbaler*/footballer (42), *fudbal*/football|soccer (40), *godina*/year (34), *liga*/ league (31), *sezona*/season (31), *rezultat*/result (29), *šansa*/chance (27), *pobeda*/victory|win (26), *trener*/coach|manager (26), *ekipa*/team-squad (25), *igra*/game/play (24), *prilika*/ opportunity (23), *situacija*/situation (23), *napadač*/striker|attacker (22).

The most common adjectives modifying *utakmica*/match are: *prva*/first (978), *kvalifikaciona*/qualifying (663), *prijateljska*/friendly (514), *poslednja*/last|final (360), *važna*/ important (281), *teška*/tough|difficult (265), *velika*/big|major (252), *dobra*/good (241), *fudbalska*/football|soccer-related (219), *evropska*/European (189), *druga*/second (165), *naredna*/next|upcoming (155), *prvenstvena*/league-related|championship (136), *odigrana*/ played (132), t*akmičarska*/competitive (117), *finalna*/final|decisive (114).

The analysis of the adjective + noun pattern in the *esFudKo* corpus used a different CQL pattern due to differences in part-of-speech tagging: [espos="ADJ"] [espos="NC|NMEA|NMON"]. This extraction yielded 271 candidates with a frequency higher than 50 and 724 candidates with a frequency exceeding 20. Table 2 presents the most frequent adjective + noun collocations in *esFudKo.*

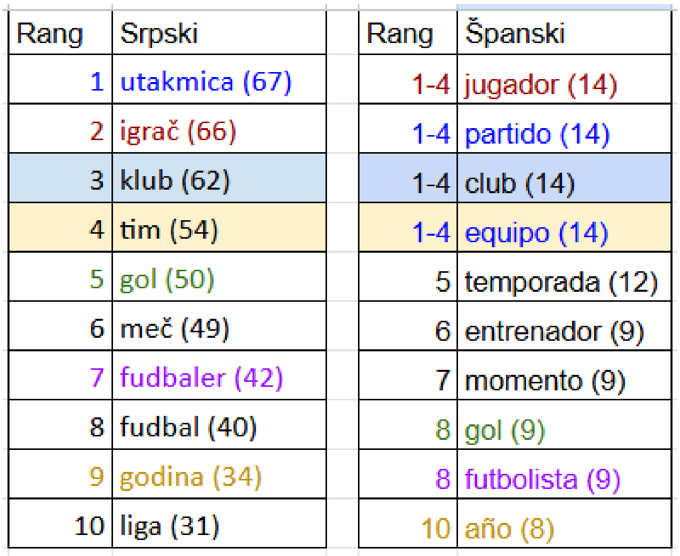
Table 2. Most frequent collocations for pattern adjective + noun in corpus esFudKo

eslemma	F	eslemma	F	eslemma	F
Real Madrid	6798	bueno|mejor versión	343	buen|mejor equipo	228
Real Sociedad	1815	bueno partido	330	cristiano Ronaldo	228
próxima temporada	1063	mucha gente	328	gran|grande partido	225
último partido	942	bueno|mejor momento	318	próximo día	221
último año	879	nuevo entrenador	310	último encuentro	221
última jornada	781	última temporada	307	gran|grande parte	214
último hora	538	buena sensación	300	buena parte	211
lateral derecho	478	última semana	297	alto nivel	210
bueno|mejor jugador	456	buen resultado	278	ex-jugador	208
último día	436	buena|mejor manera	275	media hora	206
buen jugador	435	último mes	259	gran jugador	205
último minuto	427	grande categoría	254	delantero central	203
recto final	426	mismo tiempo	254	último tiempo	200
buen momento	399	buen equipo	249		
grande|máximo goleador	359	próxima semana	239		

In the *esFudKo* corpus, for the adjective + noun pattern where the noun is preceded by an adjective, the following examples stand out based on the number of collocates: *jugador/*player (14), *partido/*match|game (14), *club/*club (14), *equipo/*team|squad (14), *temporada/*season (12), *entrenador/*coach|manager (9), *momento/*moment (9), *gol/*goal (9), *futbolista*/footballer|soccer player (9), *año/*year (8), *ocasión/*opportunity|occasion (8), *jornada/*matchday|round (8).

The adjectives characterizing *jugador/*player are: *bueno|mejor/*good|better (456), *buen*/good (435), *ex*/former (208), *grande*/great|big (205), *nuevo*/new (177), *gran|grande/* great (155), *propio*/own (91), *único*/unique (82), *mismo*/same (69), *joven*/young (63), *actual*/ current|present (29), *último*/last|latest (29), *tanto*/so much (24), *excelente/*excellent (20).

When we compare the ten most frequent nouns for Serbian and Spanish that are part of those collocations, we can notice that they overlap. In Figure 3, with the same colour and background, translation equivalents are presented. The first Serbian noun is *utakmica/*match (67), and the Spanish *jugador/*player (14) is in first place. More precisely, the first four have the same frequency in the Spanish dataset.



Rang	Srpski		Rang	Španski
1	utakmica (67)		1-4	jugador (14)
2	igrač (66)		1-4	partido (14)
3	klub (62)		1-4	club (14)
4	tim (54)		1-4	equipo (14)
5	gol (50)		5	temporada (12)
6	meč (49)		6	entrenador (9)
7	fudbaler (42)		7	momento (9)
8	fudbal (40)		8	gol (9)
9	godina (34)		8	futbolista (9)
10	liga (31)		10	año (8)

Figure 3. The most frequent nouns in the adjective + noun pattern collocation *srFuDKo* and *esFudKo*

4.3. Other Patterns with Nouns within Corpora srFudKo and esFudKo

The analysis of the Serbian corpus *srFudKo* and the Spanish corpus *esFudKo* reveals interesting differences in the use of the words *lopta/*ball and Spanish *balón*/ball, which manifest through distinct collocation types and distribution patterns. In both corpora, *lopta* and *balón* most frequently collocate with nouns related to different elements of the game. However, there are significant differences in these nouns and their functional roles.

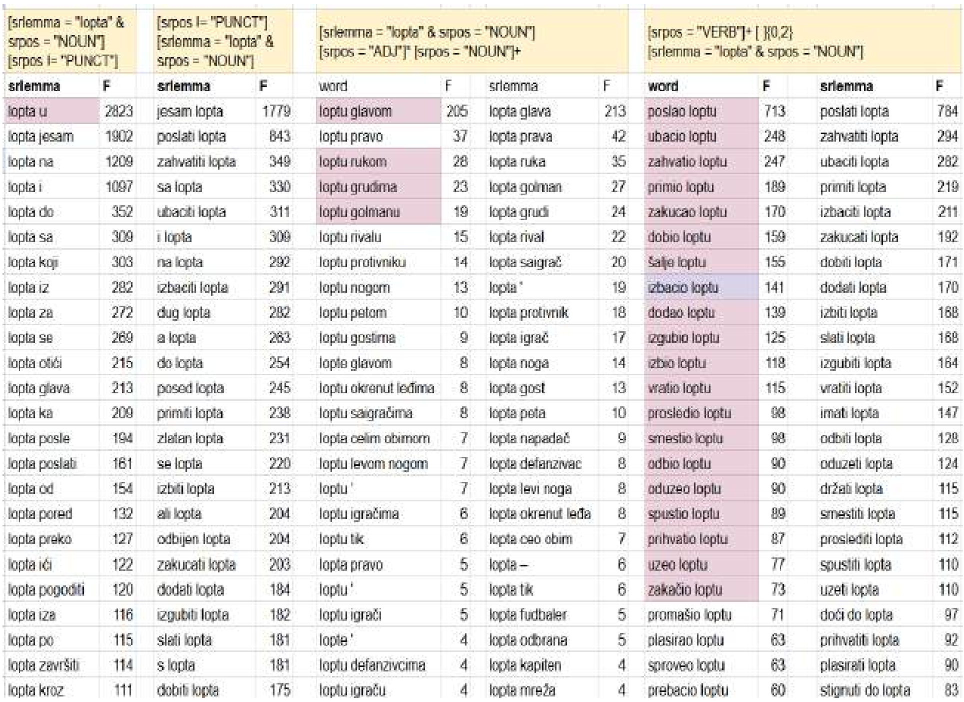
In *srFudKo*, *lopta/*ball is most commonly associated with nouns that describe physical aspects of the game and players, such as *mreža/*net, *glava/*head, *korner/*corner, *noga/*leg, *šesnaesterac/*penalty area, *golman/*goalkeeper, *udarac/*shot, *teren/*field, and *igrač/*player. These collocations clearly describe direct aspects of football and the physical actions involved, emphasizing players' interactions with the ball.

In contrast, within *esFudKo, balón/*ball frequently appears in collocations related to broader game concepts and ball possession, such as *posesión/*possession, *control/*control, *dominio/*dominance, *juego/*game|play, and *acción/*action. While there are also collocates like *pie/*foot, *área*/penalty area, and *gol*/goal, with greater focus on tactical aspects of the game suggests that the Spanish language of football emphasizes conceptual and strategic elements more than physical actions.

In the research on football-related collocations, the following queries focus on the noun *lopta/*ball, as an example, identifying its syntactic and semantic relationships with surrounding words. The first two queries, [srlemma = "lopta" & srpos = "NOUN"] [srpos != "PUNCT"] and [srpos != "PUNCT"] [srlemma = "lopta" & srpos = "NOUN"], retrieve instances where *lopta* appears before or after any word that is not punctuation, ensuring a clean extraction of collocational structures. The third query, [srlemma = "lopta" & srpos = "NOUN"] [srpos = "ADJ"]* [srpos = "NOUN"]+, captures multi-word noun phrases involving *lopta*, such as *lopta za igru*/game ball or *lopta u mreži*/ball in the net, reflecting common football terminology. The final query, [srpos = "VERB"]+ [ ]{0,2} [srlemma = "lopta" & srpos = "NOUN"], focuses on verb + noun collocations, allowing for the identification of structures like *šutirati loptu*/kick the ball, *uhvatiti loptu*/catch the ball, or *dodati loptu*/pass the ball. Figure 4 presents some results extracted from these queries that help uncover syntactic patterns and lexical preferences within Serbian football discourse, providing insights into phraseology, word order, and semantic roles in sports language.

In Spanish, the noun *balón/*ball most frequently appears with verbs such as *tocar/*touch, *parar/*stop, *robar/*steal, *mandar/*send, *perder/*lose, *controlar/*control, *rodar/*roll, and *despejar/*clear. While these verbs also refer to physical actions, there is a stronger emphasis on ball control (*controlar*, *parar*, *tocar*), confirming that the Spanish football language places greater importance on possession and tactical play compared to the Serbian football language.

A particularly interesting contrast emerges when examining the most common collocations in each language. In Serbian, *lopta/*ball is most frequently associated with the verb *poslati/*send, emphasizing a direct, action-based approach. In Spanish, however, *balón*/ball is most frequently linked with the adjective *oro*/gold, forming the syntagm *Balón de Oro*, referring to the most prestigious award for the world's best football player. It has no direct equivalent in Serbian in terms of awards, while its title is used through a literal translation – *Zlatna lopta*.



[srlemma = "lopta" & srpos = "NOUN"] [srpos != "PUNCT"]		[srpos != "PUNCT"] [srlemma = "lopta" & srpos = "NOUN"]		[srlemma = "lopta" & srpos = "NOUN"] [srpos = "ADJ"]* [srpos = "NOUN"]+				[srpos = "VERB"]+ [ ]{0,2} [srlemma = "lopta" & srpos = "NOUN"]			
srlemma	**F**	**srlemma**	**F**	word	F	srlemma	F	**word**	**F**	**srlemma**	**F**
lopta u	2823	jesam lopta	1779	loptu glavom	205	lopta glava	213	poslao loptu	713	poslati lopta	784
lopta jesam	1902	poslati lopta	843	loptu pravo	37	lopta prava	42	ubacio loptu	248	zahvatiti lopta	294
lopta na	1209	zahvatiti lopta	349	loptu rukom	28	lopta ruka	35	zahvatio loptu	247	ubaciti lopta	282
lopta i	1097	sa lopta	330	loptu grudima	23	lopta golman	27	primio loptu	189	primiti lopta	219
lopta do	352	ubaciti lopta	311	loptu golmanu	19	lopta grudi	24	zakucao loptu	170	izbaciti lopta	211
lopta sa	309	i lopta	309	loptu rivalu	15	lopta rival	22	dobio loptu	159	zakucati lopta	192
lopta koji	303	na lopta	292	loptu protivniku	14	lopta saigrač	20	šalje loptu	155	dobiti lopta	171
lopta iz	282	izbaciti lopta	291	loptu nogom	13	lopta '	19	izbacio loptu	141	dodati lopta	170
lopta za	272	dug lopta	282	loptu petom	10	lopta protivnik	18	dodao loptu	139	izbiti lopta	168
lopta se	269	a lopta	263	loptu gostima	9	lopta igrač	17	izgubio loptu	125	slati lopta	168
lopta otići	215	do lopta	254	lopte glavom	8	lopta noga	14	izbio loptu	118	izgubiti lopta	164
lopta glava	213	posed lopta	245	loptu okrenut leđima	8	lopta gost	13	vratio loptu	115	vratiti lopta	152
lopta ka	209	primiti lopta	238	loptu saigračima	8	lopta peta	10	prosledio loptu	98	imati lopta	147
lopta posle	194	zlatan lopta	231	lopta celim obimom	7	lopta napadač	9	smestio loptu	98	odbiti lopta	128
lopta poslati	161	se lopta	220	loptu levom nogom	7	lopta defanzivac	8	odbio loptu	90	oduzeti lopta	124
lopta od	154	izbiti lopta	213	loptu '	7	lopta levi noga	8	oduzeo loptu	90	držati lopta	115
lopta pored	132	ali lopta	204	loptu igračima	6	lopta okrenut leđa	8	spustio loptu	89	smestiti lopta	115
lopta preko	127	odbijen lopta	204	loptu tik	6	lopta ceo obim	7	prihvatio loptu	87	proslediti lopta	112
lopta ići	122	zakucati lopta	203	lopta pravo	5	lopta –	6	uzeo loptu	77	spustiti lopta	110
lopta pogoditi	120	dodati lopta	184	loptu '	5	lopta tik	6	zakačio loptu	73	uzeti lopta	110
lopta iza	116	izgubiti lopta	182	loptu igrači	5	lopta fudbaler	5	promašio loptu	71	doći do lopta	97
lopta po	115	slati lopta	181	lopte '	4	lopta odbrana	5	plasirao loptu	63	prihvatiti lopta	92
lopta završiti	114	s lopta	181	loptu defanzivcima	4	lopta kapiten	4	sproveo loptu	63	plasirati lopta	90
lopta kroz	111	dobiti lopta	175	loptu igraču	4	lopta mreža	4	prebacio loptu	60	stignuti do lopta	83

Figure 4. The examples of multiword terms retrieved from *srFudKo*

4.4. Beyond Frequency: Context and Semantic Nuances in Football Collocations

Sports fans incorporate sports-related expressions into everyday speech, gradually introducing them to the wider public. These phrases often preserve their original meaning as they transition from the football field, through journalists and fans, to common usage (Lavrić 2008). In addition to analyzing the frequency of occurrence, this study examines other factors such as semantic proximity and context of use. Some collocations in football terminology carry specific meanings or appear only in particular situations, reflecting tactical nuances and player strategies.

This is shown in the following examples:

1) *dati gol iz mrtvog ugla/*score a goal from a dead angle represents that this phrase is used in situations where a player scores from an unexpected or difficult position, often from an angle where it seems nearly impossible for the ball to enter the goal;

2) *izvući penal/*draw a penalty– while the referee can award a penalty kick. The phrase *izvući penal* implies that the player skillfully exploited a situation or initiated contact with an opponent in a way that

convinced the referee to award a penalty (this highlights the tactical and psychological aspects of the game);

3) *igrati bunker/*play bunker defense – a defensive strategy where a team withdraws into a compact defensive formation, avoiding offensive play and waiting for a counter-attack opportunity (the primary goal is to protect their own net, often used against stronger opponents or when defending a lead). These collocations demonstrate how football language extends beyond the literal meanings, incorporating strategic, physical, and tactical elements, making them essential for understanding football discourse in Serbian.

We also examine whether previously identified collocations are understandable to the general public, particularly to those who do not watch football and are unfamiliar with football-specific language. If a speaker from the general public can understand and use these phrases outside a sports context, it indicates that the collocations have transcended their original football domain, becoming part of the public domain. For example:

1) *igrati na sigurno/*play it safe in football. This phrase refers to a strategy of avoiding risks, such as a team focusing on defense rather than aggressive attacks, while in everyday language, it describes a cautious approach to decision-making: "I'd rather play it safe and invest in low-risk funds.";

2) *autogol/*own goal. This term is originally a football term describing a situation where a player accidentally scores against their team. It is now widely used to describe self-inflicted mistakes in any context: "…by rejecting that job offer, he scored an own goal".

3) *igrati na domaćem terenu/*to play on home ground (at home). In football, this denotes a match played in the team's own stadium or home venue, where the team typically has an advantage due to familiarity with the pitch, support from local fans, and absence of travel fatigue. In the public domain, it means that someone is in a familiar, favorable environment, where they have an advantage: ".. The president insisted the meeting be held in Belgrade because he wants to play on home ground." with the meaning: using a favorable location to gain strategic advantage in negotiations.

These examples illustrate how deeply ingrained football terminology has become within the General speech, showcasing the way in which the language of sport, especially football, influences the broader cultural and linguistic expression.

5. Conclusion

The language of sport in both Serbian and Spanish is predominantly shaped by football terminology, which has transcended its original domain, becoming an integral part of the General public discourse. Certain football-related terms are equally prevalent in both languages, reflecting the global reach and cultural significance of football. The analysis of the noun *utakmica/*match and its collocations highlight the lexical-semantic connections within the language of football. Understanding these patterns enhances linguistic research, sports journalism, and NLP applications related to sports language. Comparing Serbian and Spanish football terminology further deepens insights into how football-related words function across languages. This study contributes to these insights by analyzing the connections between collocations and multi-word football terms, which is particularly important when multi-word terms contain collocates with a clear meaning within the footballing domain. This research showcases the frequency of collocations and their possible combinations, highlighting recurring patterns that define football discourse in both languages.

Acknowledgment

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

References

[1] Heiden, S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. 24th Pacific Asia conference on language, information and computation, 2(3), 389–398.

[2] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. Lexicography, 1(1), 7–36.

[3] Krstev, C., Stanković, R., Obradović, I., & Lazic, B. (2015). Terminology Acquisition and Description Using Lexical Resources and Local Grammars. TIA, 81–89.

[4] Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I., & Kolonja, L. (2021). A data-driven approach for raw material terminology. Applied Sciences, 11(7), 2892.

[5] Lavric, E. (2008). The linguistics of football (Sv. 38). BoD–Books on Demand.

[6] Lazarević, J. V. (2024). Tekstometrija u korpusnoj linvistici: srpski fudbalski korpus *srFudKo*. Nasleđe Kragujevac, XXI(59), 111–124. https://doi.org/10.46793/NasKg2459.111L

[7] Lazarević, J., & Kitanović, O. (2024). Building the Dictionary of Football Terminology Through Data-Driven and Ontolex Model. Infotheca - Journal for Digital Humanities, 24(1), pp. 29–52, DOI 10.18485/infotheca.2024.24.1.2.

[8] Lazarević, J., Stanković, R., Škorić, M., & Rujević, B. (2023). Football terminology: Compilation and transformation into OntoLex-Lemon resource. U S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, & A. Salgado (Ur.), Proceedings of the 4th Conference on Language, Data and Knowledge (str. 634–645). NOVA CLUNL, Portugal. https://aclanthology.org/2023.ldk-1.69

[9] Levin, M. (2008). Hitting the back of the net just before the final whistle: High-frequency phrases in football reporting. The linguistics of football, 143–155.

[10] Pejović, A. (2007). Frazeologija u jeziku španskih novina. Nasleđe, 4(8), 155-167.

[11] Pejović, A. (2015). Kontrastivna frazeologija španskog i srpskog jezika. Filološko-umetnički fakultet.

[12] Prćić, T. (2016). Semantika i pragmatika reči. Filozofski fakultet.

[13] Stanković, R., Sandrih, B., Krstev, C., Utvić, M., & Škorić, M. (2020). Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for Serbian. LREC 2020-12th International Conference on Language Resources and Evaluation, Conference Proceedings, 3954–3962.

[14] Stanković, R., Krstev, C., Todorović, B. Š., & Škorić, M. (2021). Annotation of the Serbian ELTeC Collection. Infotheca–Journal for Digital Humanities, 21(2), 43-59.

Kontrastivna analiza sintaksičkih obrazaca u komparabilnim korpusima fudbala na španskom i srpskom jeziku

Jelena Lazarević, Olivera Kitanović

Sažetak

Ovaj rad istražuje pojam kolokabilnosti kao načina na koji se leksičke jedinice kombinuju sa različitim kategorijama reči kako bi formirale veće jedinice, korišćenjem analize semantičkih i sintaksičkih principa srpskih i španskih fudbalskih termina korišćenjem uporedivih korpusa: *srFudKo* i *esFudKo*, razvijenih u okviru doktorske disertacije Jelene Lazarević: *Jezičke odlike diskursa novih medija o fudbalu: kontrastivna analiza na srpskom i španskom jeziku*. Korpus *srFudKo* obuhvata 10.100.553 tokena sa pet srpskih veb-sajtova sa vestima iz fudbala, dok se *esFudKo* sastoji od 9.106.812 tokena sa dva španska fudbalska veb-sajta. Oba korpusa se analiziraju korišćenjem metoda korpusne lingvistike korišćenjem CQL (Corpus Query Language) i tekstometrijske analize. Studija ispituje kolokacije na osnovu frekvencije i semantičke privlačnosti, identifikujući različite tipove, kao što su: pridev + imenica, imenica + imenica, glagol + imenica, prilog + pridev, glagol + predloška fraza i glagol + prilog. Proces ekstrakcije kolokacije oslanja se na sintaksičke obrasce i analizu frekvencija. Osim frekvencije, istraživanje razmatra semantičku bliskost, kao i kontekst, procenjujući da li kolokacije nose specifična značenja i da li su razumljive široj javnosti. Kada se široko shvate, kolokacije iz posmatranih korpusa prevazilaze fudbalski domen. Ovo istraživanje takođe ispituje veze između kolokacija i višečlanihnih termina, pokazujući kako termini specifični za fudbal doprinose širem razumevanju sportske jezičke strukture u srpskom i španskom jeziku.

Ključne reči: fudbal, korpusi, terminologija, kolokacije, srpski, španski