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

Jelena Lazarević, Olivera Kitanović

Contrastive Analysis of Syntax Patterns in Comparable Football Corpora in Spanish and Serbian Languages | Jelena Lazarević, Olivera Kitanović | South Slavic Languages in the Digital Environment JuDig Book of Abstracts, University of Belgrade - Faculty of Philology, Serbia, November 21-23, 2024 | 2024. | |

http://dr.rgf.bg.ac.rs/s/repo/item/0009141

Jelena Lazarević
Univerzitet u Beogradu, Filološki fakultet, doktorand
E-mail: jelazarevic1@gmail.com

Olivera Kitanović
Univerzitet u Beogradu, Rudarsko-geološki fakultet
E-mail: olivera.kitanovic@rgf.bg.ac.rs

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

The aim of the paper is to explore collocability as a manner in which lexical units are combined with words from different categories, forming larger units. The research of the semantic and syntactic principles of these combinations of Spanish and Serbian footballing terms was carried out on the comparable football corpora *SrFudKo* and *EsFudko* 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 *SrFudKo* was developed through texts about football from five Serbian web news sites: *B92, Blic, Mondo, Politika,* and *Sport klub*, containing 10,100,553 tokens, of which 8,618,426 words. The corpus of Spanish-language texts on football *EsFudKo*, comes from two Spanish sites: *Marca fútbol* and *Mundo deportivo*, containing 9,106,812 tokens, of which 8,024,164 words. Both corpora to which corpus linguistics methods have been applied for data extraction are located on the platform https://noske.jerteh.rs, 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 chance. We will 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). Collocation extraction represents a technique in Computational linguistics that identifies collocations in 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, 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 the previously identified collocations are understandable to the general public who do not follow sports and are not versed in the language of football. If a speaker from the general understands them, then the collocations have surpassed their origin in the football domain, becoming part of the public domain.

The contribution of the research also means analyzing the connections between collocations and multi-part terms. Their connection is strong when the multi-part terms contain collocates that have a clear meaning within the domain of football. 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.

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

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

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(048)(0.034.2)
004.8(048)(0.034.2)

INTERNATIONAL Conference South Slavic Languages in the Digital Environment JuDig (2024 ; Beograd)

Book of abstracts [Elektronski izvor] / International Conference South Slavic Languages in the Digital Environment JuDig, International Conference South Slavic Languages in the Digital Environment JuDig, November 21-23. 2024, [Belgrade] ; [organisers University of Belgrade - Faculty of Philology [and] Society for Language Resources and Technologies JeRTeh]. - Belgrade : University, Faculty of Philology, 2024 (Belgrade : University, Faculty of Philology). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 100.

ISBN 978-86-6153-754-7

а) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157072905