The Many Faces of SrpKor

Duško Vitas, Ranka Stanković, Cvetana Krstev

Duško Vitas, Ranka Stanković, Cvetana Krstev
Society for language resources and technologies JeRTeh
E-mail: {vitas|cvetana|ranka}@jerteh.rs

The Many Faces of SrpKor

The acronym SrpKor denotes a family of electronic corpora of the modern Serbian language, the construction of which began at the end of the seventies of the last century, and which became more widely visible to the interested research community with the publication of its first version on the web in 2002. In this long period, especially before the emergence of useful textual resources on the web, corpus development consisted of the collection and processing of material as well as the development of corpus processing methods. Namely, an electronic corpus is not only a collection of texts in digital form (as, for example, it is stated in (Dobrić 2012)), but includes several components that will make such a collection useful in linguistic and other research. These components, in addition to the texts themselves, constitute, above all, software support for the organization and exploitation of the collection of texts and means for different levels of annotation of the texts that will be found in the corpus (Ви-тас 2023).

SrpKor, taking into account these components, underwent various metamorphoses during its construction, which provide a picture of the evolution of software support for the construction and exploitation of corpora, as well as the development of annotation systems at different levels (meta-data, morphological marking, lemmatization, named entities, etc.).

Extremely modest conditions (compared to other environments, both in the number of researchers involved in the construction of the corpus, allocated financial resources from different sources, and available equipment) imposed a strategy of gradual development of the corpus, which implied that new versions of the corpus would rely on material prepared and used in those versions that preceded it.

The paper will illustrate the evolution in the development of SrpKor from its first version until today, following the influx of different resources used in the construction of individual versions, as well as the changes in dimensions and text annotation system. The structure of the individual versions of the corpus, their dimensions, the period covered, and the level of annotation will be described in particular.

The basic ideas when conceiving the corpus are first presented in (Vitas, Popović 2023), and then in (Utvić 2013), where numerous details for the 2013 version of SrKor are described. Corpus interactions with dictionaries are discussed in (Krstev Vitas 2005), (Vitas, Krstev 2012).

It is important to note that texts in the Serbian language from parallelized corpora that were created at the same time as SrpKor were also included in SrpKor. In this way, the influence of web content on the composition of the corpus was partially compensated. On the other hand, such texts, which are, as a rule, extremely significant in the cultural sense, not only are not present in the material from the web but are not even included in traditional lexicographic corpora. They consist of selected scientific, literary, philosophical, anthropological, historical, and similar texts taken from reputable editions.

Further work on the development of this corpus will include, on the one hand, the enrichment of metadata, the addition of annotations and the introduction of new content. Enrichment of metadata will enable the creation of subcorpus according to different dimensions: by pronunciation, period, and domain, in addition to the ones available so far by author, register, and years. Along with the division into sentences and the addition of annotations with named

entities everywhere, the plan is to enrich them with grammatical information. The introduction of new content expands the time dimension by preparing novels, travelogues, memoirs, and historical newspapers that are valuable not only from a linguistic but also from a cultural-historical point of view, with the usual addition of (selected) content from the web.

The coupling of the Leximirka lexical base and the SrpKor corpus family is two-way, from the Leximirka interface, direct insight into examples of word use in context or in syntactic patterns is possible (Lazić, Škorić 2020). The system for lemmatization is improved from version to version, in which the Serbian morphological dictionaries play a special role, which, using the Unitex system, ensures the generation of all inflectional forms of words.

Keywords: *SrpKor, corpora, Serbian, lemmatization, Leximirka*

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

References

[1] Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005.

[2] Dobrić, Nikola. "Savremeni jezički korpusi na Zapadnom Balkanu–Istorijat, trenutno stanje i budućnost (Language Corpora in the West Balkans–History, Current State and Future Perspective)." Slavistična revija 60 (2012): 677-692.

[3] Душко Витас, Љубомир Поповић, „Конспект за изградњу референтног корпуса српског стандардног језика“, Научни састанак слависта у Вукове дане 31/1 - МСЦ, Београд, 2003, стр. 221 - 227.

[4] Душко Витас, Белешке о ручној и аутоматској обради српског језика, Језик Данас, бр. 22, 2023, Матица Српска, Нови Сад.

[5] Duško Vitas, Cvetana Krstev “Tvorbeni obrasci u elektronskom rečniku srpskog jezika”, Međunarodni komitet slavista. Komisija za tvorbu reči. Međunarodna naučna konferencija Tvorba reči i njeni resursi u slovenskim jezicima (14), pp. 515-525, 2012, Filološki fakultet Univerziteta u Beogradu, Beograd, ISBN 978-86-6153-116-3

[6] Utvić, Miloš. 2013. “Izgradnja referentnog korpusa savremenog srpskog jezika.” PhD diss., Univerzitet u Beogradu, Filoloˇski fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091

[7] Biljana Lazić, Mihailo Škorić. "From DELA Based Dictionary to Leximirka Lexical Database" in Infotheca, Faculty of Philology, University of Belgrade (2020). https://doi.org/10.18485/infotheca.2019.19.2.4

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

a) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157072905