Slovníky ve světě moderních technologií

Dictionaries in the World of Modern Technology

Workshop

Praha 27.-28. 5. 2026

Abstrakty

Abstracts

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava

Optimizing the Sketch (Engine) Grammar for a Large-Scale Dictionary Project

Keywords: lexicography, collocation analysis, Sketch Engine

Despite the advent of large language models (LLMs) into all areas of scientific research—including linguistics—we believe that the need for (manual) analysis of corpus data in the compilation of dictionary entries will not disappear in the foreseeable future.

Within our project of a new multi-volume *Dictionary of the Contemporary Slovak Language*[1] (Jarošová – Benko, 2012), our primary source of information on Slovak lexis is the gradually expanding *Slovak National Corpus* (currently *prim-11.0*[2]), supplemented by data from several web corpora (mainly *Araneum Slovacum VIII*[3]), which together amount to nearly 7 billion tokens.

Given this size, it is unrealistic to expect the lexicographers to analyze occurrences of the lexical units by simply reading through concordance lines—even words in the medium-frequency range can have tens to hundreds of thousands of occurrences in texts.[4] We thus find ourselves in the opposite situation compared to the relatively recent past—instead of a lack of data, we have to 'fight' with an abundance of it.

Probably the best tool for this 'fight' is the *Sketch Engine* corpus manager (Kilgarriff et al. 2004; 2014), which has become an indispensable daily companion for lexicographers in our project, used exactly as anticipated by its (commercial) license: each lexicographer turns it on at the start of the workday and turns it off only at the end ;-)

In addition to a reliably lemmatized and morphologically annotated corpus, effective use of the *Sketch Engine* also requires a so-called *collocation grammar* ('*sketch grammar*'), i.e., a set of rules used to generate a *collocation profile* (a '*word sketch*') consisting of tables containing the most salient collocates of the analyzed word. The majority of collocation grammars (e.g., those for corpora available on the *Sketch Engine Portal*[5]) were created according to the methodology proposed by the system's author, A. Kilgarriff. In our presentation, we introduce a different approach to creating such a grammar, motivated primarily by the needs of our dictionary project.

References:

Benko, V. (2014): Compatible Sketch Grammars for Comparable Corpora. In: *Proceedings of the XVI EURALEX International Congress: The User in Focus*: 15–19 July 2014 Bolzano/Bozen. Ed. A. Abel, Ch. Vettori, N. Rall, 417–430.

Jarošová, A. – Benko, V. (2012): The Dictionary of the Contemporary Slovak Language: A Product of Tradition and Innovation. In: *Proceedings of the 15th EURALEX International Congress.* 7–11 August 2012. Eds. R. Vatvedt Fjeld, J. M. Torjusen. Oslo 2012, 257–261.

[1] https://www.juls.savba.sk/pub_sssj.html
[2] https://korpus.sk/en/corpora-and-databases/snc-corpora/
[3] http://aranea.juls.savba.sk/aranea_about/
[4] For example, the word '*slovník*' ('dictionary') has 131,425 occurrences (19.92 i.p.m.) in the 6.6 Megatoken *Omnia Slovaca Maior* corpus.
[5] https://www.sketchengine.eu/

Kilgarriff, A. et. al. (2004): Kilgarriff, A. – Rychlý, P. – Smrž, P. – Tugwell, D.: The Sketch Engine. In: *Proceedings of the 11th EURALEX International Congress*, 105–116.
Kilgarriff, A. et al. (2014): Kilgarriff, A. – Baisa, V. – Bušta, J. – Jakubíček, M. – Kovář, V. – Michelfeit, J. – Rychlý, P. – Suchomel, V.: The Sketch Engine: ten years on. *Lexicography*, 1, 7–36.

David Blažek
Slovanský ústav AV ČR, v. v. i., Praha

K sémantickému třídění dat ve *Slovinsko-českém slovníku*: mezi definicí, ekvivalentem a kontextem

Klíčová slova: bilingvní lexikografie, sémantické třídění, sémantická anotace, Slovinsko-český slovník, umělá inteligence

Příspěvek se věnuje problematice sémantického třídění lexikálních dat ve *Slovinsko-českém slovníku* a zaměřuje se na vztah mezi definicí, překladovým ekvivalentem a kontextovými informacemi z hlediska jejich využitelnosti pro takové třídění. V centru pozornosti stojí otázka, jakým způsobem lze jednotlivé významy lexikálních jednotek systematicky zařazovat do širších sémantických kategorií a jak tyto kategorie využít při organizaci a vyhledávání slovníkových dat.

Pozornost bude věnována mj. možnostem využití umělé inteligence při sémantické anotaci slovníkových hesel. Příspěvek zároveň představí návrh pracovního postupu kombinujícího automatické metody s následnou lexikografickou validací.

Součástí příspěvku bude rovněž úvaha o možnostech využití sémanticky strukturovaných dat při konverzi stávajícího *Slovinsko-českého slovníku* do česko-slovinského slovníku.

Katja Brankačkec, Jana Kocková, Karolína Skwarska, Božana Niševa, David Blažek
Slovanský ústav AV ČR, v. v. i., Praha

Když si lexikograf povídá s počítačem: metodické výzvy a otázky při generování ilustrací pro překladové slovníky

Klíčová slova: překladový slovník, vizualizace, lexikografie, prototyp, stereotyp, prompty

Příspěvek představuje náš začínající projekt, jehož cílem je vytvoření moderní podoby digitálního překladového slovníku doplněnou o vizualizace některých významů. K vytváření vizualizací je vyvíjena aplikace založená na umělé inteligenci. Náš příspěvek tak otevírá řadu otázek, se kterými se lexikograf na této cestě setká: která hesla potřebují vizualizaci, kde je to smysluplné? Jak najít správnou vizuální reprezentaci svého pojmu (prototyp)? Rozumí AI prototypům, nebo je třeba, aby prototypičnost určil člověk? Má se vizualizace týkat výchozího jazyka, cílového jazyka nebo obou jazyků? Jak stanovit míru stereotypizace, která je už nežádoucí? Jak se vypořádat s kulturně specifickými podobami určitých předmětů? A především: jak správně zadávat prompty – otázka jazyka, komplexního chápání pojmu u člověka versus výsledek od umělé inteligence, problém polysémie, nejednoznačnosti apod. V neposlední řadě se zastavíme i u otázky, jak zacházet s dostupnými předlohami v souladu s etikou a právem.

Bronislava Chocholová
Jazykovedný ústav Ľ. Štúra SAV, v. v. i., Bratislava

Bude robiť slovníky namiesto nás umelá inteligencia?

Kľúčové slová: výkladový slovník, slovníkové heslo, exemplifikácia, typické korpusové príklady

Významnou súčasťou slovníkového hesla predovšetkým v jednojazyčných výkladových slovníkoch je exemplifikácia (uvádzanie príkladov). Nazdávame sa, že táto zóna slovníkového hesla je v komplementárnom vzťahu s výkladom – nielenže dokladá existenciu slova, resp. jeho významu (významov), ukazuje aj jeho gramatické a štylisticko-pragmatické vlastnosti a môže sprostredkovať rôzne encyklopedické či kulturologické informácie (Janočková – Hašanová – Chocholová 2026).

Vo svojej prezentácii sa zameriame na niektoré aspekty spojené s exemplifikáciou vo výkladovom slovníku, konkrétne v pripravovanom *Slovníku súčasného slovenského jazyka* (SSSJ; *A – G*, 2006; *H – L*, 2011; *M – N*, 2015; *O – Pn*, 2021, na 5. zväzku sa stále pracuje). SSSJ vzniká na korpusovom základe, v súčasnosti autorsko-redaktorský kolektív pracuje s interným korpusom Omnia Slovaca IV Maior Beta (23.01), ktorý zahŕňa hlavný korpus SNK, verziu prim-9.0 a webové korpusy, pričom na vyhľadávanie v korpusoch sa využíva programový nástroj Sketch Engine, ktorý vytvára tvarový a kolokačný profil hľadanej lexémy. Slovníkové doklady sa tak preberajú priamo z korpusu, príp. sa čiastočne upravujú a modifikujú skracovaním. Výber a prípadná úprava dokladov je v rukách autorov či redaktorov, pričom stále prevláda presvedčenie, že dôležitú úlohu zohráva ich jazykové povedomie a odborné skúsenosti.

No je to naozaj tak? S využitím webovej aplikácie *Typické korpusové príklady*, ktorú vytvoril R. Garabík z JÚĽŠ (Garabík – Karčová 2025), sa pokúsime na viacerých jedno- a viacvýznamových slovníkových heslách ukázať, či toto východisko stále platí alebo sa autori slovníka budú môcť úplne spoľahnúť na metódu automatického získavania krátkych zmysluplných príkladov z textových korpusov.

Literatúra:

Janočková, N. – Hašanová, J. – Chocholová, B. (2026): Všeobecné a špecifické zásady a problémy (ne)uvádzania exemplifikácie v slovníku súčasného slovenského jazyka. *Jazykovedný časopis* 77, 1 (v tlači).

Garabík, R. – Karčová, A. (2025): Analyzing grammatical anomalies in lexical data for fun and profit. In: J. Ballagó, V. Lipp (eds.): *1st International Conference on Lexicology and Lexicography. Book of abstracts*. Budapest: ELTE Research Centre for Linguistics, 23–24.

Fabian Kaulfürst, Anja Pohončowa
Serbski institut, Budyšin / Chóśebuz

Identifikacija „nowych słowow“ a jich zapřijeće do serbskich leksikografiskich projektow

Klučowe słowa: leksikografija, monitoring pismowstwa, „nowe słowa“, korpus

Tuchwilu dźěła so w Serbskim instituće na wjacorych leksikografiskich projektach za wobě serbšćinje. Jich wuslědki su na rěčnymaj portalomaj *hornjoserbsce.de* resp. *dolnoserbski.de* přistupne. Při dźěle so systematisce za leksiku slědźi, kotraž njeje dotal w leksikaliskej datowej bance instituta za kóždu rěč registrowana. Při tym móže so wo woprawdźite neologizmy jednać, ale tež wo staršu leksiku, kotraž njeje swój puć do słownikow a z tym do datoweje banki namakała. Přednošk předstaja dwaj systematiskej přistupaj, tajku leksiku identifikować a za wozjewjenje we wšelakich leksikografiskich projektach spřihotować.

Veronika Kolářová, Václava Kettnerová, Michal Olbrich, Jiří Mírovský
Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha

NomVallex: Valence českých substantiv a adjektiv v sítích slovotvorně příbuzných slov

Klíčová slova: substantiva, adjektiva, slovotvorné vztahy, valence, valenční slovník

NomVallex je ručně anotovaný valenční slovník, který zachycuje valenci českých substantiv a adjektiv (verze 2.5 obsahuje 1337 lexikálních jednotek v celkovém počtu 730 lexémů; Kolářová et al. 2024). U substantiv i adjektiv se zaměřuje na následující slovotvorné typy: deverbální, deadjektivní, desubstantivní a primární. Slovník je koncipován jako lexikografický zdroj umožňující výzkum valence slovotvorně příbuzných lexikálních jednotek, proto v relevantních případech poskytuje odkaz od určité lexikální jednotky k odpovídající lexikální jednotce jejího základového slova, obsaženého buď v NomVallexu, např. *uraženost* < *uražený*, nebo (v případě sloves) ve slovníku VALLEX (Lopatková et al. 2022), např. *uražený* < *urazit se – urážet se*. U propojených lexikálních jednotek jsou pak automaticky porovnávány jejich valenční rámce; např. u substantiva *uraženost* jsou rozdíly oproti adjektivu *uražený* zapsány následovně: =ACT(>:↑->2,pos) =ADDR(=:na+4;+:k+3) =PAT(=:že;+:nad+7,z+2;-:pro+4).

V tomto příspěvku představíme, jak si lze zachycené slovotvorné vztahy a s nimi související rozdíly ve valenci nově zobrazit v celých sítích slovotvorně příbuzných slov, a to v uživatelsky přívětivé podobě, přímo na webových stránkách slovníku.

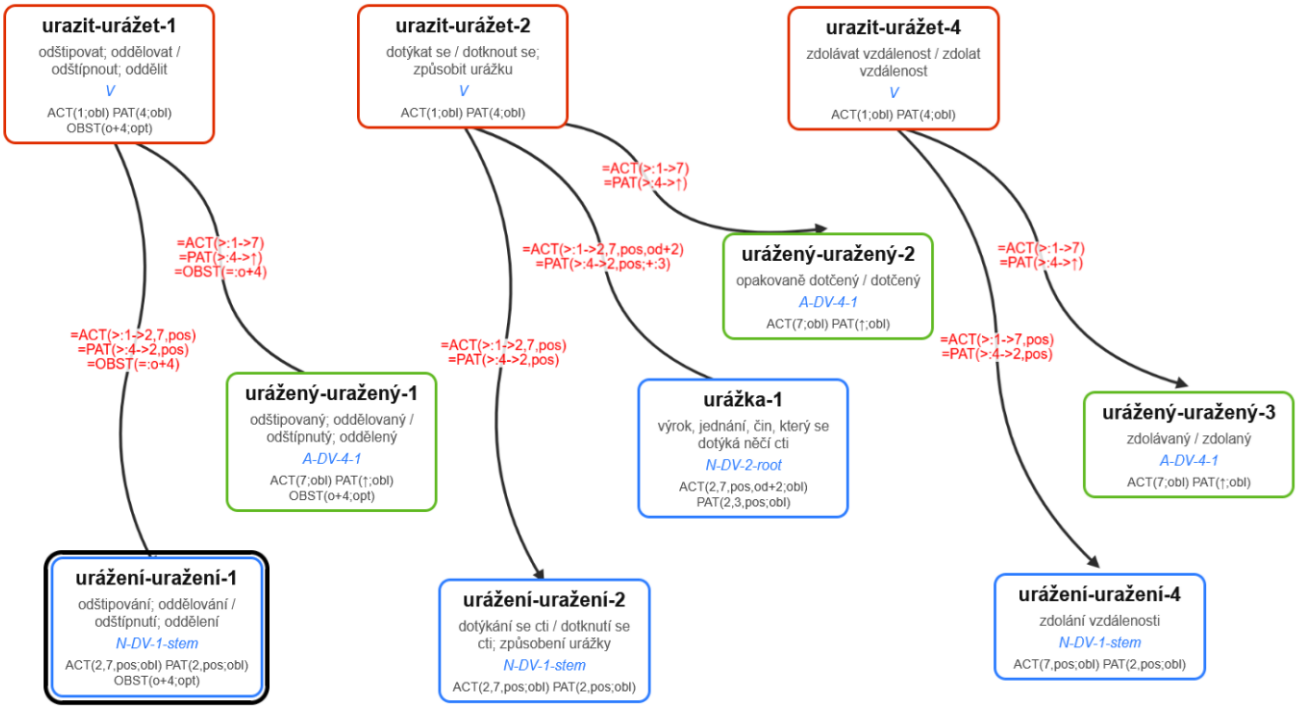
Při zobrazení slovotvorných vztahů se inspirujeme velkými lexikálními databázemi, které se zaměřují na automatické modelování derivačních vztahů, pro češtinu zejména databází DeriNet (Olbrich et al. 2025). Oproti sítím DeriNetu, kde jednotlivé uzly odpovídají lexémům (lemmatům), jsou v NomVallexu slovotvorné vztahy a rozdíly ve valenci zachycovány na úrovni lexikálních jednotek, tedy lexikálních významů. To umožňuje rozlišit, že zatímco např. substantivum *uražení* a adjektivum *uražený* mají několik významů odpovídajících slovesným významům *odštípnout*, *dotknout se* a *zdolat vzdálenost*, substantivum *urážka* odpovídá pouze významu *dotknout se* (Obr. 1). V sítích NomVallexu jsou barevně rozlišeny jednotlivé slovní druhy, přičemž každý uzel pro lexikální jednotku obsahuje parafrázi jejího významu, derivační typ a valenční rámec. Rozdíly ve valenci základového a odvozeného slova jsou připojeny k hranám mezi danými uzly (pro substantivum *uraženost* viz Obr. 2). Sítě slovotvorně příbuzných slov v NomVallexu obsahují v průměru 3 uzly, nejobsáhlejší síť zahrnuje 46 lexikálních jednotek.

Bibliografie:

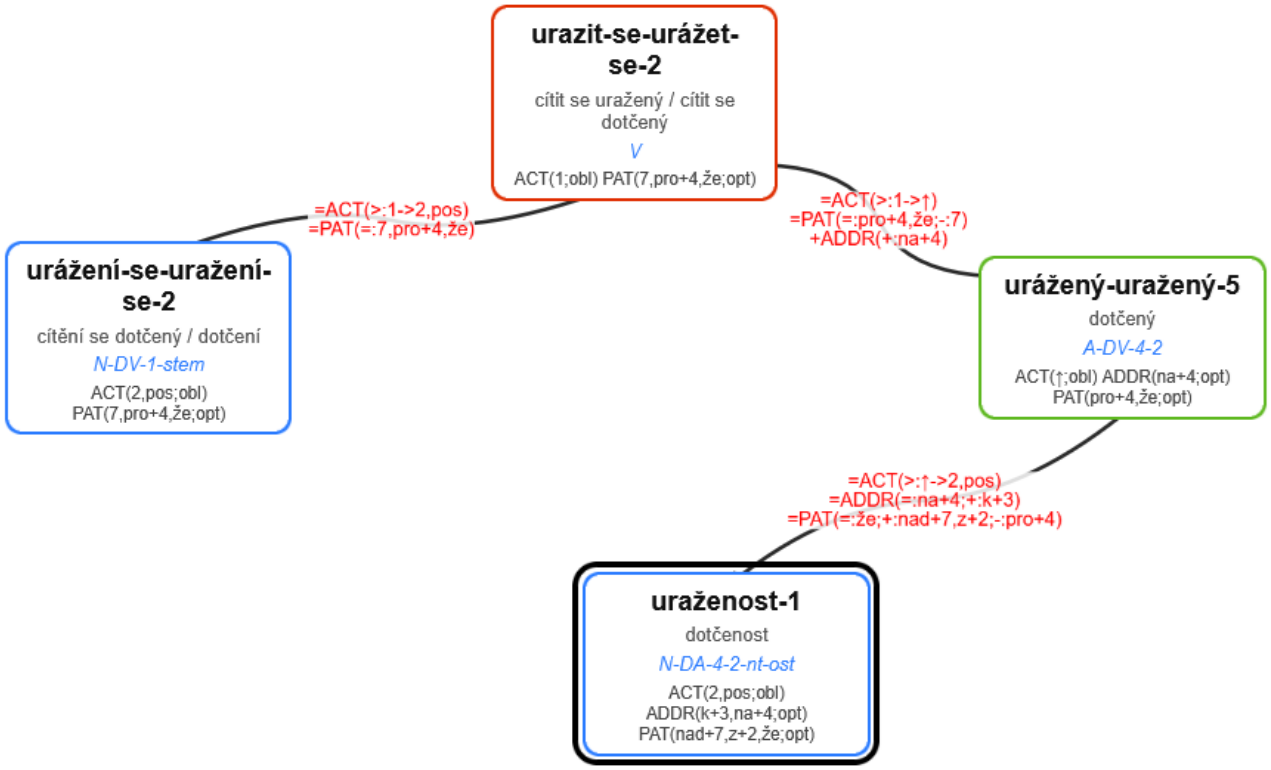
Kolářová, V. et al. (2024). *NomVallex 2.5*. LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK, http://hdl.handle.net/11234/1-5826, available on https://ufal.mff.cuni.cz/nomvallex/2.5/.

Lopatková, M. et al. (2022). *VALLEX 4.5.* LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK, http://hdl.handle.net/11234/1-4756.

Olbrich, M. et al. (2025). *DeriNet 2.3.* LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK, http://hdl.handle.net/11234/1-5846.



Obr. 1: Deriváty různých významů slovesa *urazit – urážet*, tedy 1: *odštípnout – odštipovat*, 2: *dotknout se – dotýkat se*, 4: *zdolat vzdálenost – zdolávat vzdálenost*



Obr. 2: Vizualizace slovotvorných vztahů a rozdílů ve valenci v NomVallexu (přímé a nepřímé deriváty slovesa *urazit se – urážet se*)

Ana Mihaljević, Josip Mihaljević
Old Church Slavonic Institute, Zagreb

AI in the Retrodigitization and Compilation of the Dictionary of the Croatian Redaction of Church Slavonic

Keywords: Croatian Church Slavonic, retrodigitization, historical lexicography, AI

The *Dictionary of the Croatian Redaction of Church Slavonic* has been the primary long-term research project of the Old Church Slavonic Institute in Zagreb since 1991. Printed fascicles covering the entries from A to I have been published and, since 2024, they have been undergoing a process of retrodigitization. This presentation explores the role of artificial intelligence and modern digital tools in the transformation of the printed dictionary into a digital resource and in supporting the preparation of future fascicles. A demo version of the electronic dictionary is available at https://rjecnik.stin.hr/.

The retrodigitization process involves several technical and methodological challenges. The original volumes exist primarily as legacy PDFs without editable text, which required AI-assisted optical text recognition and structural reconstruction during the creation of the digital portal. Additional complexity arises from the multilingual and multiscript nature of the dictionary: entries and examples involve the Latin script, Old Cyrillic, Glagolitic, and the Greek alphabet, and the languages included are Croatian Church Slavonic, Croatian, English, Latin, and Greek. This use of multiple different scripts makes dictionary entries much harder to write, index, search, and display on the web.

Another challenge concerns the extraction and normalization of headword lists, since the printed dictionary follows the traditional azbuka (Cyrillic alphabetical order) rather than the modern Latin alphabetical sequence. The digital version therefore required automated and semi-automated methods for identifying, organizing, and linking entries.

The paper will also demonstrate newly implemented hyperlinking possibilities within the dictionary and towards external resources. Finally, several experimental tests will be presented in which different large language models (LLMs) were evaluated for assisting lexicographic work, including the preparation of draft dictionary entry based on representative corpus examples. LLMs were also used in programming custom functions for our dictionary system, including custom search functions, entry ordering, and automatic script and font conversion (e.g. from Latin to Glagolitic, Cyrillic, or Greek based on the input field).

Special attention will be placed on the preparation of a machine-processable corpus derived from the dictionary's original handwritten card index, written mostly in Old Cyrillic and containing multiple handwriting styles. The results illustrate both the potential and the current limitations of AI in historical lexicography.

Tanja Mirtič
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana

Zvukové nahrávky ve výkladovém slovníku

Klíčová slova: zvukové nahrávky, uměla inteligence, výkladový slovník, spisovná mluva

Nový výkladový slovník současného slovinského jazyka eSSKJ, který vzniká v Ústavu pro slovinský jazyk Frana Ramovše, obsahuje také zvukové nahrávky rodilého mluvčího. Ke zvukovým nahrávkám lze přistupovat kliknutím na ikonu reproduktoru ve výchozím slovníkovém zobrazení. Každá nahrávka obsahuje základní podobu slova a první vedlejší tvar. Jelikož je dynamický přízvuk odvoditelný z melodického, nahrává se pouze výslovnost s melodickým přízvukem.

Doposud byla slova namluvena dvěma různými mluvčími. V budoucnu plánujeme tyto zvukové nahrávky přidat i do dalších příruček na slovníkovém portálu Fran, z toho důvodu jsme v roce 2025 dokončili pilotní nahrávání a konkurz nových mluvčích. Na naši výzvu k účasti v projektu reagovalo 40 mluvčích spisovné slovinštiny, kteří jsou v kontaktu se spisovným jazykem a většina z nich jej svou profesní činností také spoluutváří (hlasatelé, novináře, herci). Shromážděné nahrávky následně na základě předem sjednocené metodologie analyzovalo podle různých kritérií 15 odborníků na mluvený jazyk. Na závěr jsme vybrali několik mluvčích, kteří se budou podílet na dalším nahrávání zvukových záznamů.

Zpracování pořízených nahrávek je předmětem vícefázového procesu. Nahrávky je potřeba nastříhat do jednotlivých bloků, normalizovat a zkontrolovat kvalitu, ale zejména zkontrolovat kvalitu z pohledu výslovnosti. V příspěvku se pokusíme zjistit, do jaké míry nám může umělá inteligence v tomto procesu pomoci, zda a jak nám může usnadnit práci v jednotlivých fázích přípravy nahrávek.

Hana Mžourková, Barbora Martinkovičová, Martin Beneš, Veronika Štěpánová
Czech Language Institute of the Czech Academy of Sciences, Prague

The Internet Language Reference Book among Dictionaries

Keywords: Internet Language Reference Book, monolingual dictionary, user friendliness, and utility of language sources, consistency of language sources

The Internet Language Reference Book (*Internetová jazyková příručka*, hereafter ILRB) is probably the best-known and most widely used source of information about Czech for the general public. After its explanatory section was published online in 2008, the ILRB has been substantially expanded; its dictionary section now contains over 112,000 entries. Conceived as a user-oriented project, its content was selected and organized according to the needs of users of the language consulting service. Today, the ILRB – especially in orthography and morphology – constitutes a comprehensive online resource on Czech, replacing older academic handbooks that can no longer adequately reflect changes in the language norm.

Although the ILRB does not provide definitions of word meanings, it is not an isolated resource. Its dictionary section mediates information from older explanatory dictionaries – Dictionary of Standard Czech language (*Slovník spisovného jazyka českého*), Dictionary of Standard Czech (*Slovník spisovné češtiny*), and New Academic Dictionary of Foreign Words (*Nový akademický slovník cizích slov*) as well as from the most recent dictionary – Academic Dictionary of Contemporary Czech (*Akademický slovník současné češtiny*, hereafter ADCC). The ILRB is interconnected with other resources: selected entries link to the Language Inquiry Database (*Databáze jazykových dotazů*), and to the application Word at a Glance (*Slovo v kostce*) of the Czech National Corpus.

The content of the ILRB is continuously updated, based on research conducted by the Department of Language Cultivation, which ensures its operation and further development, and in cooperation with the Department of Contemporary Lexicology and Lexicography, which has been building the ADCC (online since 2017). This collaboration responds to the need for consistency among current sources of information about Czech produced at the Czech Language Institute, while also emphasizing user-friendliness (Bergenholtz – Bergenholtz, 2011; Lew – de Schryver, 2014) and utility (Mžourková – Křivan, 2019).

For the first time, our paper presents the activities that ensure the consistency of these reference works. As the ILRB is involved in the ongoing Lexical Portal of Czech project, we also explain the role of ILRB data within the portal and its relationship to other sources.

References:

Bergenholtz, H. – Bergenholtz, I. (2011): A dictionary is a tool, a good dictionary is a monofunctional tool. In: Pedro A. Fuertes-Olivera – Henning Bergenholtz (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London – New York, NY: Continuum.

Lew, R. – de Schryver, G.-M. (2014): Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), 341–359. <https://doi.org/10.1093/ijl/ecu011>.

Mžourková, H. – Křivan, J. (2019): Pusťme uživatele do slovníku! Aneb o neprobádané cestě v české lexikografii. *Naše řeč*, 102(1–2), 36–50.

Petya Osenova
Sofia University "St. Kl. Ohridski" and Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia

Strategies for Dictionary Creation

Keywords: dictionary creation platform, large corpora, AI, dictionary-corpus interface

The talk features two interrelated strategies for dictionary creation in Bulgarian.

The *first one* is based on an in-house dictionary development platform, called CLaDA-BG-Dict (Angelov et al. 2022). There the lexicographer has at disposal all the main instruments (create, edit, delete, observe, extract, sort, etc.) for creating and/or managing various lexical entries within the specific dictionaries. In addition, the CLaDA-BG-Dict platform provides access to various text corpora and other dictionaries (domain specific, bilingual, diachronic, etc.). The representation formats can be customized to the content type of the dictionary. Currently, the platform hosts the following dictionaries, among others: *Bulgarian Valency dictionary*, *BTB-WordNet*, *Bulgarian Hurt Lexicon*, *Bulgarian phraseological dictionary*. It also provides a linking possibility to open knowledge databases like Wikipedia, Wikidata, etc. CLaDA-BG-Dict ensures two important facilities: a) access to contexts of word and phraseological units' usages (through corpora, other dictionaries, related examples, world knowledge databases) and b) a possibility to interact with various AI instruments which have not been integrated into the platform yet. These instruments include data, extracted from the platform, as input to LLMs (e.g. the lemma paradigms are fed into an LLM, and the LLM extracts and clusters all the detected examples from a large corpus). In the near future we also plan to use AI for automatic definition generation and LLM training over synthetic corpora based on dictionaries of thesauri type like WordNets (Agirre et al. 2014).

The second strategy focuses on the combined usage of large corpora and AI for meaning detection and discrimination, definition creation, examples collection, phraseology handling (metaphorical vs. literal usages). In this case I will discuss some possible scenarios where the CLASSLA large corpus of Bulgarian Web (5 billion words)[6] is employed together with relevant prompts to ChatGPT/Gemini/Claude, etc.

Thus, both strategies – dictionary-platform-based and large-corpus-based presuppose the inclusion of AI-oriented steps. However, for the moment AI has been used as an outer tool, not as part of them in contrast to other applications that already provide this integration upon the necessary licence.[7]

The dictionary development platform has been developed within CLaDA-BG infrastructure, while the CLASSLA corpus of Bulgarian – in close cooperation with the CLARIN Knowledge Centre for South Slavic languages – CLASSLA (Ljubešić – Kuzman 2024),[8] hosted by CLARIN-Slovenia.

References

Agirre et al. 2014: Agirre, E. – López de Lacalle, O. – Soroa, A.: Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics* 40(1), 57–84.

Angelov et al. 2022: Angelov, Z. – Simov, K. – Osenova, P. – Kancheva, Z.: The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In: Erjavec, T. – Eskevich, M.

[6] https://www.clarin.si/ske/#dashboard?corpname=classlaweb2_bg
[7] https://www.english-corpora.org/ai-llms/
[8] https://www.clarin.eu/k-centres/classla

(eds): *Selected papers from the CLARIN Annual Conference 2022*. Linköping Electronic Conference Proceedings 198, 12–22. https://doi.org/10.3384/ecp198002

Ljubešić – Kuzman 2024: Ljubešić, N. – Kuzman, T.: CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* Torino, Italia. ELRA and ICCL, 3271–3282.

Simov et al. 2017: Simov, K. – Osenova, P. – Popov, A.: Comparison of Word Embeddings from Different Knowledge Graphs. In: Gracia, J. – Bond, F. – McCrae, J. – Buitelaar, P. – Chiarcos, C. – Hellmann, S. (eds.): *Language, Data, and Knowledge*. LDK 2017. Lecture Notes in Computer Science, vol 10318. Springer. https://doi.org/10.1007/978-3-319-59888-8_19

Andrej Perdih, Dejan Gabrovšek, Janoš Ježovnik
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana

Automatizované přiřazování obrazů ke slovníkovým významům

Klíčová slova: lexikografie, sémantika, obrazy, přiřazování obrazů k významům, rozpoznávání obrazu

Obrazy ve slovnících představují vytoužený, ale pro náročné zpracování a pro velké množství základních údajů neprioritní obsah slovníků (Biesaga, 2016; 2017a; 2017b; Dziemianko, 2022; Lišková – Šemelík, 2024). Pro uživatele obrazy představují užitečné doplnění heslové statě v roli znázorňování slovníkových významů grafickým způsobem jako pomůcka k lepšímu nebo rychlejšímu porozumění výkladu významů.

Příspěvek přináší výsledky analýzy přiřazování obrazů k významům ve *Slovníku slovinského spisovného jazyka* (druhé vydání, 2014) pomocí dávkového zpracování jazykovým modelem GPT-4o mini se schopností rozpoznávání obrazů (Perdih et al., 2025). Naše východisko představuje databáze pedagogického portálu slovinského jazyka *Franček* (Gabrovšek et al. 2026; Perdih et al., 2021; 2024), ve které 64 063 obrazů bylo manuálně přiřazeno k 19 760 slovníkovým heslům, primárně však bez významových informací. V rámci následného přiřazování obrazů ke slovníkovým významům byla ověřena míra spolehlivosti přiřazování jazykovým modelem.

Vybraných 398 polysémních podstatných jmen s 1 572 obrazy bylo anotováno lidským anotátorem a jazykovým modelem GPT-4o mini s výstupním omezením 300 tokenů. Vstupní informace byly podány ve formátu JSONL. Výzva obsahovala heslové slovo, URL obrazu, instrukce a výklady významů včetně ID čísla významu v databázi. Jazykový model vykázal relativně vysokou celkovou shodu s lidským anotátorem (tj. 85,1 %, Cohenova kappa (κ) 0,70). Míra shody byla vyšší, když jazykový model hodnotil pouze odpovídající významy, a nižší, když hodnotil také neodpovídající významy v rámci hesla. Pro kontrolu spolehlivosti lidského anotátora byli do aktivity zapojeni další dva lidští anotátoři, výsledky ale i přes neúplnou shodu mezi nimi ukazují stabilní míru shody s jazykovým modelem (85,7 % a 86,9 % všech jednotek, Cohenova kappa 0,71 a 0,73).

Literatura:

Biesaga, M. (2016): Pictorial Illustration in Dictionaries. The State of Theoretical Art. In: T. Margalitadze – G. Meladze (Ed.): *Proceedings of the 17th EURALEX International Congress*. Ivane Javakhishvili Tbilisi University Press, 99–108. <http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_007_p99.pdf>.

Biesaga, M. (2017a): Dictionary Tradition vs. Pictorial Corpora: Which Vocabulary Thematic Fields Should Be Illustrated? *Lexikos* 27, 132–151. https://doi.org/10.5788/27-1-1397

Biesaga, M. (2017b): Pictorial Illustrations in Encyclopaedias and in Dictionaries – a Comparison. In: I. Kosem – C. Tiberius – M. Jakubíček – J. Kallas – S. Krek – V. Baisa (Ed.): *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, 221–236. Lexical Computing CZ. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper13.pdf>.

Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL* 34(2), 218–234. https://doi.org/10.1017/S0958344021000264

Franček. <https://www.franček.si/>.

Gabrovšek et al. (2026): Gabrovšek, D. – Ježovnik, J. – Pavlič, M. – Perdih, A. (2026): Metodologija priprave nabora slik za pedagoški portal Franček: izbor, semantična

opredelitev in nadaljnja uporaba. *Jezikoslovni zapiski* 32(1), 53–68. https://doi.org/10.3986/JZ.32.1.03

Lišková, M. – Šemelík, M. (2024). Show me the meaning of being lonely... Graphic Illustrations in The Academic Dictionary of Contemporary Czech. In: K. Š. Despot – A. Ostroški Anić – I. Brač (Eds.): *Lexicography and Semantics. Book of Abstracts of the XXXI EURALEX International Congress*. Institut za hrvatski jezik, 163–165. <https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex_boa_17.pdf>.

Perdih et al. (2024): Perdih, A. – Ahačič, K. – Jakop, N. – Ledinek, N. – Petric Žižić, Š.: Semantic Information on the Franček Educational Language Portal for Slovenian. In: K. Š. Despot – A. Ostroški Anić – I. Brač (Eds.): *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institut za hrvatski jezik, 155–168. <https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralax-XXI-final-web.pdf>.

Perdih et al. (2021): Perdih, A. – Ahačič, K. – Ježovnik, J. – Race, D.: Building an Educational Language Portal Using Existing Dictionary Data. *Journal of Linguistics/Jazykovedný Časopis*, 72(2), 568–578. https://doi.org/10.2478/jazcas-2021-0052

Perdih et al. (2025): Perdih, A. – Gabrovšek, D. – Ježovnik, J.: Image-to-sense alignment using AI tools. In: I. Kosem – M. Jakubíček – M. Medveď – K. Zgaga – Š. Arhar Holdt – T. Munda – A. Salgado (Eds.): *e-Lex 2025*, 861–874. <https://elex.link/elex2025/wp-content/uploads/elex2025_proceedings.pdf>.

Ivan P. Petrov
Department of Slavonic Studies, University of Vienna

The Construction of Bilingual Greek–Old Church Slavonic Dictionaries and the Challenges of Digital Lexicography

Keywords: digital lexicography; digital philology; Old Church Slavonic; bilingual lexicography; corpus-based lexicography; Greek–Slavonic translation

This paper presents the conceptual framework, methodology, and results of the project *The Vocabulary of Constantine of Preslav's Didactic Gospel: Old Bulgarian–Greek and Greek–Old Bulgarian Word Lists*, funded by the Bulgarian National Science Fund and conducted at the Institute of Balkan Studies and Centre for Thracology. Drawing on selected examples, the presentation outlines the prerequisites, structure, and methodological challenges underlying the digital tools developed within the project.

Constantine of Preslav's *Didactic Gospel* represents the earliest and most substantial Slavic homiletic corpus, translated into Old Church Slavonic from Byzantine Greek sources, including Gospel catecheses and homilies. The collection comprises fifty-one homilies corresponding to the Sundays of the liturgical year, supplemented by original authorial contributions in the prefaces and postfaces. Despite its importance for early Slavic literary culture, the text became fully accessible to scholarship only with the complete edition published in 2012 by Maria Tichova.

By integrating the traditions of classical Palaeoslavonic lexicography with approaches from digital philology, the project produced both publishable lexical indices and adaptable digital tools applicable to other corpora. The development of these resources required addressing several methodological challenges, including the fluidity of the textual base, discrepancies between Greek and Slavic lexicographical traditions, and the problem of asymmetrical translation relationships. These issues, along with their practical implications for bilingual dictionary construction, form the core of the discussion.

Selected literature:

Rabus, A. (2019): Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus. *Scripta & E-Scripta* 19 (2019): 9–32.

Ruskov, M. – Taseva, L. (2022): Computer-Aided Modelling of the Bilingual Word Indices to the Ninth-Century Uchitel'noe Evangelie. In: *Proceedings of 1st International Workshop on Digital Platforms and Resources for Access to Literary Heritage* (DIPRAL), 19–30 <https://doi.org/10.48550/arXiv.2211.05579>.

Tihova, M. (2012): *Starobălgarskoto Učitelno evangelie na Konstantin Preslavski*. Monumenta Linguae Slavicae Dialecti Veteris. Fontes et Dissertartiones 58. Freiburg im Breisgau.

Veronika Štěpánová, Barbora Martinkovičová
Department of Language Cultivation, Czech Language Institute, Czech Academy of Sciences, Prague

Orthoepy in the World of Modern Technology

Keywords: orthoepy (standard pronunciation), Czech, AI, speech recognition, lexicography

The paper focuses on the practical use of smart technologies, particularly tools based on artificial intelligence, in research on the pronunciation of contemporary Czech. It presents the methodology of current orthoepic analyses, which are primarily employed for lexicographic and language-consulting purposes, and from a historical perspective recalls how orthoepic research on Czech was conducted in previous decades and what its limitations were in comparison with present-day possibilities. Before the widespread adoption of AI-based technologies, it was not possible to ensure fast and reliable speech-to-text conversion in sufficiently large datasets. Compared to corpus research on written language, adequate phonetic analysis was therefore considerably more complex and, in some cases, practically unfeasible. A fundamental shift in orthoepic research has been brought about by the development of new technologies. One of the newly available options is to use the robust media monitoring database *Newton One*, where spoken language is automatically recognized and transcribed with high accuracy. This significantly facilitates the collection of authentic evidence of the phonetic realization of the units under investigation, particularly words and sound sequences. Relevant occurrences can be quickly searched for in orthographic transcripts, and the corresponding segments of recordings can be played back immediately. In this way, it is now possible to obtain a sufficient number of examples (ideally several dozen) from different speakers, especially professional speakers (so-called model speakers, who are considered carriers of the norm; cf. the concept of language standardization in Ammon, 1995), on the basis of which a normative pronunciation form can be identified. This form can then be appropriately reflected in pronunciation recommendations in dictionaries (currently the *Academic Dictionary of Contemporary Czech*) and in other language sources, such as the *Internet Language Reference Book* and the *Language Inquiry Database*.

Sources:

Academic Dictionary of Contemporary Czech = Akademický slovník současné češtiny (2012–2026) [on-line]. Praha: Ústav pro jazyk český AV ČR <www.slovnikcestiny.cz>.

Ammon, U. (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz*. Berlin – New York: Walter de Gruyter.

Internet Language Reference Book = Internetová jazyková příručka (2008–2026) [on-line]. Praha: Ústav pro jazyk český AV ČR <https://prirucka.ujc.cas.cz>.

Language Inquiry Database = Databáze jazykových dotazů (2016–2026) [on-line]. Praha: Ústav pro jazyk český AV ČR <https://dotazy.ujc.cas.cz/>.

Newton One: Media Database. (2026) [on-line]. Praha: Newton Media, a. s., <https://app.newtonmedia.eu/cs-cz/>.

Pavel Vondřička, Michal Škrabal
Ústav lingvistiky, Filozofická fakulta Univerzity Karlovy, Praha

Od lístečků k online slovníku: Proces rozšiřování torza Mudrova slovníku

Klíčová slova: horní lužická srbština, čeština, TshwaneLex, Elasticsearch, Hotko2, Hornjoserbski tekstowy korpus, InterCorp

Aktuálně vznikající hornolužicko-český slovník Mudra 2.0 je zpracováván pomocí lexikografického softwaru TshwaneLex a fulltextového vyhledávače Elasticsearch. V příspěvku popíšeme proces zpracování jednotlivých hesel od původních pramenů (textové zdroje, kartotéční lístky s excerpty), přes editaci a doplňování těchto zdrojů ve specializované databázi až po finální zveřejnění hesel v pravidelných aktualizacích. Důležitým pramenem, který neměl Jiří Mudra k dispozici, jsou korpusová data, pocházející ať už z jednojazyčných (Hotko2, Hornjoserbski tekstowy korpus) nebo paralelních (InterCorp) korpusů.

Daniel Zoba, Beata Brězanowa, Anita Hendrichowa, Kryštof Peršín
Załožba za serbski lud, Rěčny centrum Witaj, Budyšin

Challenges in Automatic Translation from Upper Sorbian to German

Keywords: speech recognition, machine translation, whisper, language model, AI

In this paper, the challenges for development of a good automatic translation from Upper Sorbian to German (and possibly other languages) will be demonstrated and discussed.

The system for simultaneous translation is composed of an Upper Sorbian speech recognition system, combined with the machine translation system “so.tra”. A basic introduction of the systems for speech recognition and machine translation will be provided. It can be demonstrated that, although the overall recognition quality is good, the resulting translation still has major issues, up to the point that the translation result is hard to relate to the original utterance.

Furthermore, it will be shown how colloquial speech is handled at the recognition stage already, to prevent additional issues from machine translation.

The paper will conclude with additional challenges faced, especially on the interface between recognition and translation. Topics include voice activity detection, recognition hallucinations, domain and context, speaker ductus, incomplete sentences, and improper grammar and formulations. An outlook is given on possible compensations on either recognition or translation stage.

References:

“so.tra” machine translation: https://sotra.app/
“so.tra” backends: https://github.com/WitajSotra/modele
OpenAI whisper speech recognition model for Upper Sorbian:
https://huggingface.co/Korla/whisper-large-v3-turbo-hsb-0
META wav2vec2-bert speech recognition model for Upper Sorbian
https://huggingface.co/Korla/Wav2Vec2BertForCTC-hsb
Fraunhofer recIKTS speech recognition system:
https://www.ikts.fraunhofer.de/de/abteilungen/elektronik_mikrosystem_biomedizintechnik/pruef_analysesysteme/kognitive_materialdiagnostik.html