

Najnovšie sa začína používať aj ženská podoba *liberka*, v množnom čísle *liberky*, napr. *Žilina nastúpila bez liberky Olejníkovej, ktorú na tomto poste zastúpila smečiarka Bombeková* (SME, 2010). – *Najskúsenejšia hráčka slovenských volejbalistiek je liberka Simona Kohútová* (Šport, 2009). – *To, že máme dobrú nahrávačku a solidne blokárky i liberku, vieme už dávnejšie* (Šport, 2007). V SSSJ H – L uvádzaná správne utvorená plurálová podoba *liberovia* sa v praxi používa zriedkavo.

Ivor Ripka
Bratislava

Webový korpus slovenčiny ARANEUM + HPLT + FineWeb2

Úvod

Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied (ďalej JÚĽŠ SAV) si už dlhodobo udržiava korpusovú tradíciu v tvorbe textových korpusov a vo výskume v oblasti korpusovej lingvistiky. Tvorbe korpusov sa venuje najmä oddelenie Slovenského národného korpusu (ďalej SNK), no mnohé korpusy (vrátane tohto, ktorý opisujeme) vznikajú aj mimo neho. V sérii článkov si podrobnejšie predstavíme niektoré významnejšie korpusy, ktoré vznikli na pôde JÚĽŠ SAV a ktoré sú dostupné širokej odbornej aj laickej verejnosti. Začneme (v čase písania tohto článku) najväčším webovým korpusom slovenčiny (po odstránení duplicít), čiže aj najväčším korpusom všeobecného jazyka.

Webové korpusy zohrávajú v súčasnej korpusovej lingvistike významnú úlohu – dokonca sa stali de facto štandardom pri trénovaní veľkých jazykových modelov (z čisto pragmatických dôvodov: pre bežné písané jazyky sú webové korpusy pomerne ľahko dostupné alebo vyrobiteľné v dostatočných veľkostiach). Hoci kvalita textu (v zmysle pravopisnej a gramatickej správnosti a štylistickej uhladenosti) v týchto korpusoch často nedosahuje úroveň beletrie alebo staršej žurnalistiky, na druhej strane oveľa lepšie odrážajú „bežný jazyk“ než veľké reprezentatívne („národné“) korpusy založené na beletrii a iných publikovaných textoch.

Opisovaný webový korpus vznikol zjednotením korpusu *Araneum Slovacum VII Maximum* (Benko, 2024), slovenskej časti datasetu *High Performance Language Technologies*, verzia 2.0 *cleaned* (de Gibert et al., 2024) a slovenskej časti datasetu *Hugging Face FineWeb 2* (Penedo et al., 2024). Celý korpus bol následne deduplikovaný na úrovni odsekov, čo efektívne zabezpečilo aj odstránenie duplícit medzi týmito tromi zdrojmi.

Korpus je lematizovaný a morfológicky anotovaný nástrojom *morpho-dita* (Garabík – Mitana, 2023), jeho veľkosť je 10,4 miliardy tokenov, 8,6 miliardy slov, 33,6 milióna dokumentov. Texty v korpuse boli zbierané v rokoch 2012 až 2024 (podotýkame, že to nie sú dátumy vzniku textov; vzhľadom na charakter WWW sa čas vzniku nedá určiť, čo trochu obmedzuje využitie korpusu napr. pri štúdiu neologizmov). Korpus je prístupný prostredníctvom korpusového manažéra NoSketch Engine na adrese <https://www.juls.savba.sk/webskahfcorp.html>.

1. Štruktúra korpusu

Každému tokenu v texte prislúchajú atribúty podľa tabuľky 1. Podľa zaužívaných zvyklostí v iných korpusoch je *word* pôvodný tvar slova, tak ako sa vyskytuje v texte (vrátane zachovania veľkosti písmen), a *lc* je slovo, v ktorom sú všetky písmená konvertované z veľkých na malé. Morfológická značka je pridelená podľa slovenského morfológického tagsetu¹, používaného takmer univerzálne aj v iných slovenských korpusoch JÚLŠ SAV.

Tabuľka 1 – Atribúty tokenov

atribút		priklad
word	slovo (pôvodná veľkosť písmen)	Jazykovedným, Bratislave, jazykovedné
lemma	základný tvar slova	jazykovedný, Bratislava
tag	morfológická značka	SSmsl
lc	slovo (malé písmená)	jazykovedným, bratislave, jazykovedné

Korpus je členený do dokumentov, každý dokument zodpovedá jednému HTML dokumentu – webovej stránke. Dokumenty sú ďalej delené na

¹ <https://korpus.sk/korpusy-a-databazy/databazy/morfologicka-databaza/>

odseky <p> a vety <s> podobne ako v ostatných korpusoch JÚLŠ SAV². V tabuľke 2 uvádzame všetky štruktúry korpusu.

Z nich *doc.collection* je označenie súboru, z ktorého dokument pochádza. V prípade korpusu *Araneum Slovacum VII Maximum* je to reťazec *aranea*; v prípade datasetov HPLT a FineWeb2 korpusu má tvar *hplt:X* alebo *fineweb:X*, kde X je označenie súboru dokumentov podľa daného datasetu. *Doc.collection* tak nadobúda tieto hodnoty:

- *aranea*
- *hplt:X*
 - Pre dáta pochádzajúce z Internet Archive³ je X jedno z *wide5*, *wide6*, *wide10*, *wide11*, *wide12*, *wide15*, *wide16*, *wide17*, *survey3*, *archivebot* a označuje príslušnú zbierku z databázy Internet Archive.
 - Pre dáta pochádzajúce z Common Craw⁴ je X jedno z *cc13*, *cc14*, *cc15*, *cc16*, *cc17*, *cc18*, *cc19*, *cc20*, *cc21*, *cc22*, *cc23*. Dvojčíslia označujú dvojčíslicie roku, v ktorom boli dokumenty získané z WWW.
- *fineweb:X*
 - X má tvar CC-MAIN-yyyy-mm, kde yyyy je rok a mm mesiac získania dokumentu z WWW.

Hodnota *doc.url* odkazuje na URL stránky (v čase stiahnutia; nie nevyhnutne v súčasnosti ešte fungujúce). Z tejto hodnoty sú potom odvodené *doc.urldomain*, *doc.website* a *doc.tld*.

Vzťah medzi nimi si najlepšie ozrejníme na príklade. Napríklad stránka slovníkového portálu JÚLŠ bude mať *doc.url*="https://slovník.juls.savba.sk/", *doc.urldomain*="savba.sk", *doc.website*="slovník.juls.savba.sk", *doc.tld*="sk".

Tabuľka 2 – Štruktúry korpusu

doc	dokument (webová stránka)
doc.id	jedinečný identifikátor dokumentu
doc.collection	zdroj (aranea, fineweb alebo hplt)

² <https://www.juls.savba.sk/tools.html#Korpusy>

³ <https://archive.org>

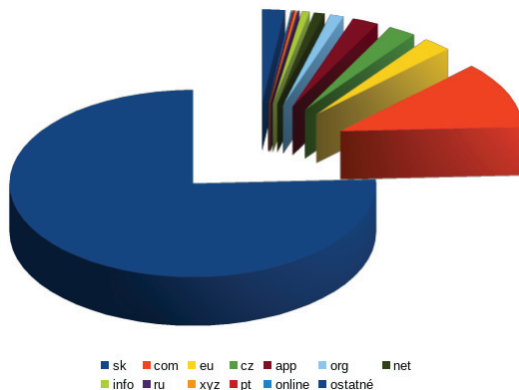
⁴ <https://commoncrawl.org>

doc.url	URL stránky
doc.urldomain	domény dvoch najvyšších úrovní
doc.website	názov servera (hostname)
doc.tld	doména najvyššej úrovne URL dokumentu
doc.date	dátum dokumentu
doc.year	rok dokumentu (automaticky získaný z dátumu)
doc.tokcount	počet tokenov v dokumente
p	odsek
s	veta
g	na tomto mieste sa medzi tokenmi nenachádza medzera

Uvedieme si pre zaujímavosť 12 najčastejších domén najvyššej úrovne podľa počtu dokumentov (webstránok).

Tabuľka 2 & Obrázok 1 – Najčastejšie domény 1. úrovne v korpuse podľa počtu dokumentov

<i>TLD</i>	<i>počet dokumentov</i>
sk	25496061
com	3888888
eu	857778
cz	837006
app	786449
org	399450
net	314445
info	223595
ru	59468
xyz	46153
pt	43027
online	33939
<i>ostatné</i>	680488



2. Iné dostupné veľké webové korpusy slovenčiny

Na stránkach JÚLŠ SAV sa môžeme stretnúť aj s ďalšími všeobecnými webovými korpusmi:

Korpusy ARANEA⁵ sú spoločným projektom JÚLŠ SAV a UNESCO Katedry plurilingválnej a multikultúrnej komunikácie Univerzity Komenského v Bratislave (Benko, 2024). Ide o rodinu webových korpusov, aktuálne v 27 jazykoch, niekedy aj s teritoriálnymi variantmi. Najväčší slovenský korpus *Araneum Slovacum VII Maximum* má veľkosť 5,3 miliardy tokenov; z neho sú vytvorené menšie podkorpusy *Araneum Slovacum VII Minus* veľkosti 125 miliónov tokenov a *Araneum Slovacum VII Maius* veľkosti 1,25 miliardy tokenov.

Korpus *web-7.0*⁶ oddelenia SNK je založený na dátach beta verzie korpusu *Araneum Slovacum VII Maximum* (24.02) z projektu Aranea. Segmentácia viet a tokenizácia sú prebraté z pôvodného korpusu *Araneum Slovacum*, ale lematizácia a morfológická anotácia je totožná s korpusom *prim-10.0*⁷.

Webový korpus slovenčiny HPLT⁸ bol spracovaný z datasetu projektu High Performance Language Technologies, verzie 2.0 cleaned (a je teda podmnožinou korpusu opísaného v tomto článku). Korpus je lematizovaný a morfológicky anotovaný, je dodatočne deduplikovaný na úrovni dokumentov, veľkosť korpusu je 5,2 miliardy tokenov, 4,3 miliardy slov, 21 miliónov dokumentov. Deduplikovaný (na úrovni odsekov) podkorpus je navyše anotovaný syntakticky. Veľkosť podkorpusu je 2,7 miliardy tokenov, 2,2 miliardy slov. Anotácie uzlov (syntaktické relácie) sú označené podľa koncepcie Universal Dependencies (de Marneffe et al. 2021).

Zhrnutie

Vytvorenie tohto korpusu poskytne verejnosti nový, významný zdroj lingvistických dát slovenského jazyka. V súčasnosti ide o najväčší dostupný (hoci aktuálne len prostredníctvom webového rozhrania) webový korpus slovenského jazyka. Je to zároveň zrejme jeden z posledných korpusov, pri ktorých si môžeme byť istí, že prevažná väčšina textov v ňom bola vytvore-

⁵ <http://aranea.juls.savba.sk>

⁶ <https://korpus.sk/webovy-korpus/>

⁷ <https://korpus.sk/korpusy-a-databazy/korpusy-snk/predchadzajuce-verzie-snk/>

⁸ <https://www.juls.savba.sk/hpltskcorp.html>

ná ľuďmi. V budúcnosti očakávame, že čoraz viac webového obsahu bude generovaného prostredníctvom veľkých jazykových modelov alebo ako strojový preklad z iných jazykov.

Korpus je prístupný na adrese <https://www.juls.savba.sk/webskahfcorp.html> prostredníctvom korpusového manažéra NoSketch Engine.

Radovan Garabík
Jazykovedný ústav Ľ. Štúra SAV, v. v. i.

LITERATÚRA

BENKO, Vladimír: The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. In: Lecture Notes in Computer Science : Text, Speech, and Dialogue. Proceedings, Part 1. - Heidelberg : Springer, 2024, zv. 15048, s. 55 – 70.

DE GIBERT, Ona – NAIL, Graeme – AREFYEV, Nikolay – BAÑÓN, Marta – VAN DER LINDE, Jelmer – JI, Shaoxiong – ZARAGOZA-BERNABEU, Jaume – AULAMO, Mikko – RAMÍREZ-SÁNCHEZ, Gema – KUTUZOV, Andrey – PYYYSALO, Sampo – OEPEN, Stephan – TIEDEMANN, Jörg: A New Massive Multilingual Dataset for High-Performance Language Technologies. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia. ELRA and ICCL. s. 1 116 – 1 128.

GARABÍK, Radovan – MITANA, Denis. Analysing Accuracy of Slovak Language Lemmatization and MSD Tagging. In: Slovenská reč, 2023, roč. 88, č. 2, s. 129 – 140.

DE MARNEFFE, Marie-Catherine – MANNING, Christopher – NIVRE, Joakim – ZEMAN, Daniel: Universal Dependencies. Computational Linguistics, 2021, roč. 47. č. 2, s. 255 – 308.

PENEDO, Guilherme – KYDLÍČEK, Hynek – SABOLČEC, Vinko – MESSMER, Bettina – FOROUTAN, Negar – JAGGI, Martin – VON WERRA, Leandro – WOLF, Thomas: FineWeb2: A sparkling update with 1000s of languages, december 2024. Dostupné z URL: <https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>.