Korpus textov rusínskej wikipédie

Úvod

Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied (ďalej JÚĽŠ) 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, no veľa korpusov (vrátane tu opisovaného) vzniká aj mimo tohto oddelenia.

V príspevku prezentujeme nový korpus textov rusínskej wikipédie[1], ktorý vznikol na pôde JÚĽŠ a ktorý sprístupňujeme prostredníctvom korpusového manažéra NoSketch Engine, čo umožňuje využívať bohaté štatistické prostriedky a možnosti korpusového manažéra na lingvistický výskum.

Rusínčina sa všeobecne považuje za ohrozený jazyk (Plišková, 2017); v Atlas of the World's Languages of Danger 2010 je klasifikovaná ako *vulnerable* – zraniteľná, s neustále klesajúcim počtom hovoriacich (Moseley, 2010). Odhady počtu používateľov jazyka sa pohybujú medzi 70-tisíc (podľa sčítania obyvateľstva) až 600-tisíc (neoficiálne odhady zahrňujúce hovoriacich na západe Ukrajiny, ktorí svoj regiolekt považujú za nárečie ukrajinčiny). Najmä používanie rusínčiny ako písaného jazyka je pomerne obmedzené. Používanie a rozvoj jazyka negatívne ovplyvňuje aj nedostatok vyučovacích materiálov – absentujú učebnice, slovníky, gramatiky.

Digitálne zdroje jazyka a nástroje (klasického) počítačového spracovania rusínčiny neexistujú prakticky vôbec. Rozvoju týchto nástrojov bráni známa roztrieštenosť literárneho jazyka: existujú dve hlavné kodifikácie[2] písanej rusínčiny používajúce cyriliku – poľská (lemkovská) (Фонтаньскій – Хомяк, 2000) a slovenská (Ябур – Панько, 1994; Панько, 1994), okrem nich aj niekoľko iných variantov (Kushko, 2007). Ďalej existuje štandardizovaný zápis latinkou, používaný na Slovensku v niektorých periodikách. Oficiálne Ukrajina rusínčinu neuznáva ako samostatný jazyk a jej používanie nepodporuje (čo je relevantné hlavne v Zakarpatskej oblasti).

---

[1] Kvôli prehľadnosti budeme v článku používať termín *wikipédia* v uvedenej podobe, aj keď v niektorých prípadoch vykazuje známky vlastného mena ako názov konkrétnej implementácie alebo konkrétnej jazykovej mutácie. Rovnaký termín budeme používať aj na označenie rusínskej jazykovej mutácie, hoci striktne vzaté by sme jej názov *Википедія* mali transliterovať.

[2] Kodifikácia v širokom ponímaní, nie nevyhnutne v zákonodarnom zmysle.

V rusínskej wikipédii[3] sa používajú oba spomínané kodifikované varianty jazyka, ako aj neoficiálny ad hoc vytvorený ďalší, vo wikipédii pomerne populárny, ktorý sa obracia skôr k historickým zdrojom jazyka, čím je na prvý pohľad odlíšiteľný (napr. používaním písmena ѣ).

V článku sa vyhneme úvahám o panónskej (vojvodinskej) rusínčine a jej klasifikácii ako semištandardizovaného variantu východoslovenčiny, písaného cyrilikou.[4]

Štruktúra korpusu

Každému tokenu v texte prislúchajú atribúty podľa tabuľky 1. Vzhľadom na neexistenciu lematizátora pre rusínčinu je viditeľná absencia často používaného atribútu lemma aj absencia atribútu opisujúceho morfologické vlastnosti slova. Označenie *word* je podľa zaužívaných zvyklostí v iných korpusoch pôvodný tvar slova, tak ako sa vyskytuje v texte (vrátane zachovania veľkosti písmen), *lc* je slovo, v ktorom sú všetky písmená konvertované z veľkých na malé, *trans* je transliterácia atribútu *lc* do latinky (používané sú iba ASCII znaky).

Tabuľka 1. Atribúty tokenov

		príklad
word	slovo (cyrilika, pôvodná veľkosť písmen)	*Руснакох*
lc	slovo (cyrilika, malé písmená)	*руснакох*
trans	slovo (ASCII transliterácia, malé písmená)	*rusnakox*

Korpus je členený do dokumentov, každý dokument zodpovedá jednému článku vo wikipédii. Dokumenty sú ďalej delené na odseky <p> a vety <s> podobne ako v ostatných (slovenskojazyčných) korpusoch JÚĽŠ[5]. Segmentácia na vety sa robila automaticky, heuristikou založenou na interpunkcii a veľkosti písmen. Šablóny úmyselne nie sú nahradzované, ale vzhľadom na ich extralingvistický charakter sa v korpuse vyskytujú ako samostatné tokeny.

[3] https://rue.wikipedia.org/

[4] Genetickú príslušnosť tohto jazyka k inej vetve reflektuje aj rusínska wikipédia, ktorá obsahuje texty výlučne v karpatskej rusínčine, a samostatná wikipédia v panónskej rusínčine existuje v inkubátore (https://incubator.wikimedia.org/wiki/Wp/rsk).

[5] https://www.juls.savba.sk/tools.html#Korpusy

Tabuľka 2. Štruktúry korpusu

doc	dokument (článok wikipédie)
doc.id	jedinečný identifikátor dokumentu
doc.url	URL stránky
doc.title	názov článku
doc.timestamp	čas poslednej editácie článku
p	odsek
s	veta
g	na tomto mieste sa nenachádza medzera

Transliterácia

Aby sme uľahčili používanie korpusu pomocou klávesníc, ktoré nepodporujú cyriliku alebo diakritiku, používame vlastnú transliteráciu, založenú zhruba na romanizácii BGN/PCGN 2016 v atribúte *trans* (v CQL hľadaniach) a tiež v „Jednoduchom vyhľadávaní“. Transliterácia nesleduje žiadny iný cieľ ani sa nesnaží o bijektívne zobrazenie medzi cyrilikou a latinkou. Používame iba znaky repertoáru ASCII, ich prehľad je v tabuľke 3.

Tabuľka 3. Transliterácia v atribúte *trans*

а	a
б	b
в	v
г	h
ґ	g
д	d

е	e
ё	jo
є	je
ж	zh
з	z
и	y

і	i
ї	ji
й	j
к	k
л	l
м	m

н	n
о	o
п	p
р	r
с	s
т	t

у	u
ф	f
х	x
ц	c
ч	ch
ш	sh

щ	sc
ъ	'
ы	y
ь	'
ѣ	ji
э	e

ю	ju
я	ja

Iné korpusy rusínskeho jazyka

Jediný nám známy podobný korpus rusínskeho jazyka vytvorený na Universität Leipzig (Leipzig Corpora Collection, 2021) tiež z textov wikipédie je staršieho dátumu ako náš korpus a poskytuje len základné štatistické

údaje potrebné na korpusový lingvistický výskum. Wikipédia (vo všetkých jazykových mutáciách) je spracovaná v podobe textových dát v datasete Plaintext wikipedia dumps (Rosa, 2018). Na rozdiel od nášho prístupu tieto texty expandujú šablóny. Okrem toho na Universität Freiburg[6] vzniká korpus hovorenej rusínčiny *Korpus des gesprochenen Russinischen*, ktorý je v štádiu rozpracovanosti.

Zhrnutie

Vytvorenie korpusu rusínskych textov je z našej strany základom k modernému lingvistickému výskumu jazyka – pre „väčšie" jazyky (ku ktorým v tomto kontexte patrí aj slovenčina) je existencia textového korpusu nevyhnutným predpokladom výskumu a rozvoja nástrojov na spracovanie jazyka. Hoci taký stav počítačového spracovania rusínčiny, aký poznáme z iných jazykov, je v nedohľadne, existencia korpusu a jeho dostupnosť je významným krokom k podpore jazyka a jeho písomnej formy a k ďalšiemu výskumu.

V čase písania tohto článku korpus obsahoval texty rusínskej wikipédie k 20. 1. 2024; korpus plánujeme priebežne, hoci nepravidelne aktualizovať. Aktuálna veľkosť korpusu je 1 274 378 tokenov, 898 454 slov, 114 252 viet, 74 884 odsekov a 9 500 dokumentov. Text je tokenizovaný a segmentovaný na vety. Pravopis sa drží originálneho pravopisu článkov vo wikipédii, vďaka čomu do značnej miery odráža pravopisné preferencie autorov článkov.

Korpus je prístupný na adrese https://www.juls.savba.sk/ruecorp.html, dostupné je vyhľadávanie prostredníctvom korpusového manažéra NoSketch Engine, ako aj celý korpus vo vertikálnom formáte.

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

LITERATÚRA

BGN/PCGN 2016: Guidance on the US Board on Geographic Names (BGN)/Permanent Committee on Geographical Names (PCGN) romanization systems. https://assets.publishing.service.gov.uk/media/5d94678d40f0b65e5cf93f3b/ROMANIZATION_OF_RUSYN.pdf [cit. 10. 1. 2024].

[6] http://www.russinisch.uni-freiburg.de/corpus

ЯБУР, Василь – ПАНЬКО, Юрій: Правила русиньского правопису. Русиньска оброда, Пряшів 1994.

ПАНЬКО, Юрій і кол.: Орфоґрафічный словник русиньского языка. Русиньска оброда, Пряшів 1994.

PLIŠKOVÁ, Anna: The mother tongue of Rusyns in the Slovak Republic after 1989: Status, problems, and perspectives. In: Approaches to Rusyn 2017. Boudovskaia, Elena (ed) Sapporo : The Slavic-Eurasian research center Hokkaido University, 2021, 1-40.

ФОНТАНЬСКІЙ, Ґ., ХОМЯК, М.: Ґраматыка лемківского языка. Katowice: Śląsk, 2000.

KUSHKO, Nadiya: Literary Standards of the Rusyn Language: The Historical Context and Contemporary Situation. In: The Slavic and East European Journal, vol. 51, no. 1, 2007, pp. 111–32.

MOSELEY, Christopher (ed): Atlas of the World's Languages in Danger. Unesco, 2010. https://unesdoc.unesco.org/ark:/48223/pf0000187026 [cit. 10. 1. 2024].

Leipzig Corpora Collection: Rusyn Wikipedia corpus based on material from 2021. Leipzig Corpora Collection. Dataset. https://corpora.uni-leipzig.de?corpusId=rue_wikipedia_2021. [cit. 10. 1. 2024].

ROSA, Rudolf: Plaintext Wikipedia dump 2018, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2735 [cit. 10. 1. 2024].