Koľko slov má slovenčina...

Ďalšia fáza budovania Národného korpusu slovenského jazyka II.

V predchádzajúcom čísle Správ SAV sme predstavili doterajšie aktivity a aktuálny stav Slovenského národného korpusu, v ďalšej časti sa sústredíme na najbližšie plány, najmä na tvorbu Slovenskej terminologickej databázy.

Jednou z najčastejších otázok, s ktorými sa v oddelení Slovenského národného korpusu JÚĽŠ SAV, ale aj pri bežnom styku s verejnosťou stretávame, je: Koľko slov má slovenčina a ktoré sú najviac a najmenej frekventované slová. Na prostrednú časť otázky sa odpovedá najjednoduchšie. Zoznam najfrekventovanejších slov a lem sa sprístupňuje spolu s novou verziou korpusu na internete (http://korpus.juls.savba.sk/stats/). Na prvých miestach sa okrem bodky a čiarky stabilne nachádzajú predložky v, na, spojka a, polyfunkčné slovko sa a pomocné sloveso byť. Kým pri najčastejších výskytoch sa hodnoty jednotlivých slov nachádzajú v dosť veľkých odstupoch od seba, v strede tabuľky majú rovnakú frekvenciu dve, tri i viaceré slová, a na konci sú slová/tvary s výskytom 3, 2 i 1, ktorých je však na týchto posledných miestach aj niekoľko tisíc s rovnakou frekvenciou. Preto sa nedá jednoznačne povedať, ktoré slovo má najnižšiu frekvenciu. Je ich veľa a okrem rôznych nesprávnych zápisov či preklepov sú to zväčša cudzie alebo archaické slová, úzko špecializované termíny a pod. Ich ojedinelý až unikátny výskyt v korpuse môže byť spôsobený nedostatkom textov z istej oblasti a po rozšírení korpusu sa môže zmeniť. Pri dobrej koncepcii budovania korpusu a prihliadaní na vyváženosť či reprezentatívnosť (primerané zastúpenie rôznorodých textov) by však okrajové javy nikdy nemali mať extrémne vysokú frekvenciu, ale ich jestvovanie by malo byť dostatočne zachytené.

Najnáročnejšia je odpoveď na prvú časť otázky o počte slov v slovenčine, ale týka sa to akéhokoľvek jazyka. V učebniciach či slovníkoch cudzích jazykov sa dočítame, že na základné porozumenie a komunikáciu stačí zvládnuť tisíc až tri tisíc slov daného jazyka. Najnovšie vydanie Krátkeho slovníka slovenského jazyka spracúva približne 60 tisíc heslových slov tvoriacich jadro slovnej zásoby, Slovník slovenského jazyka z r. 1959 – 1968 obsahoval necelých 130 tisíc hesiel a v novom veľkom výkladovom slovníku slovenčiny sa plánuje spracovať okolo 240 tisíc heslových slov. To však ešte stále nie sú všetky slová, ktoré tvoria slovenčinu, ešte sú tu nárečia, historická slovná zásoba, termíny... S veľkou pravdepodobnosťou by sme presiahli číslo milión, a to značne. Ak sme uviedli, že Slovenský národný korpus obsahuje v poslednej verzii zo začiatku tohto roka 350 miliónov textových jednotiek, ide o celkový počet slov textu v databáze korpusu (súčet výskytov rôznych slov a tvarov), pričom počet pribúdajúcich nových unikátnych slov klesá priamo úmerne s veľkosťou korpusu. Korpusové štatistiky svetových jazykov uvádzajú, že v korpuse, ktorý obsahuje 100 miliónov textových jednotiek, čo je v súčasnosti minimálna veľkosť všeobecného korpusu, sa 8 tisíc jednotiek nachádza v 95 percentách textu a zvyšných 5 percent reprezentuje 500 tisíc jednotiek. S narastaním korpusu sa tento pomer mení – počet jednotiek v 95 percentách sa zväčšuje, počet jednotiek vo zvyšných 5 percentách postupne klesá. Trend v budovaní korpusov je jednoznačný: čím viac dát, tým sú to lepšie dáta.

Slovenský národný korpus by sa mal v druhej etape svojho budovania rozrásť na 600 miliónov textových jednotiek v základnom korpuse písaných textov od r. 1955 do súčasnosti, t. j. do r.

2011, dokedy bola schválená druhá časť projektu. Tento materiál sa bude využívať predovšetkým pri tvorbe 8-zväzkového Slovníka súčasného slovenského jazyka, ktorého koncipovanie potrvá minimálne do r. 2013, ale plánuje sa na ňom aj príprava ďalších slovníkov, ktoré prispejú k exaktnejšiemu poznaniu jazykového systému slovenčiny a typologickým i iným výskumom: frekvenčný, retrográdny, kolokačný slovník. Priebežne sa bude pokračovať v sprístupňovaní lingvistických zdrojov a slovníkov na internete, pripravuje sa ich spracovanie a distribúcia aj na CD/DVD nosičoch rovnako ako aj časti hlavného korpusu. Rozširovanie paralelných korpusov sa zameria najmä na česko-slovenský a slovensko-český paralelný korpus, ktorý poslúži ako materiálová báza na tvorbu prekladového slovníka. Osobitnou súčasťou bude budovanie korpusu hovorenej slovenčiny, ktorej výskum sa stáva jednou z dôležitých úloh viacerých slovakistických pracovísk (Prešov, Banská Bystrica, Nitra, Bratislava), a tvorba Slovenskej terminologickej databázy (https://data.juls.savba.sk/std/), ktorá v súčasnosti obsahuje vyše 3 000 termínov z rôznych oblastí a mala by sa rozšíriť najmä o terminológiu z oblasti práva a ekonómie.

Napĺňanie Slovenskej terminologickej databázy a plnenie jej funkcie je determinované vo veľkej časti rozsahom a kvalitou špecializovaného podkorpusu odborných textov, z ktorých sa dajú termíny automatizovane extrahovať pomocou osobitných metód a nástrojov. Poskytovateľmi odborných textov sú viaceré mimoakademické pracoviská a vydavateľstvá, napr. Ekonomická univerzita v Bratislave, Filozofická fakulta UK v Bratislave, Slovenská poľnohospodárska univerzita v Nitre, Strojnícka fakulta Žilinskej univerzity v Žiline, Slovenské národné múzeum, Slovenská národná galéria, IURA EDITION, Jaga group, Slovenský filmový ústav, Múzeum pedagogiky. Nadväzovanie kontaktov s akademickými pracoviskami či jednotlivými pracovníkmi SAV prebieha vo viacerých fázach, zatiaľ sú medzi zazmluvnenými poskytovateľmi (http://korpus.juls.savba.sk/contributors/) napr. Ústav politických vied SAV (predtým Politologický kabinet SAV), Ústav štátu a práva SAV, Ústav etnológie SAV, prof. Š. Luby, ako aj viacerí pracovníci Ústavu orientalistiky SAV či Kabinetu divadla a filmu SAV. Spolupráca sa rozbehla s Ústavom slovenskej literatúry SAV, Astronomickým ústavom SAV a Predsedníctvom SAV – pracovisko Slovenského národného korpusu JÚĽŠ SAV digitalizuje staršie tlače, napr. Správy SAV, ktoré sa potom stávajú jednak súčasťou korpusu, jednak poskytovateľská inštitúcia získa na základe dohody elektronické verzie, ktoré použije podľa svojich potrieb. Vedeckých prác s dobre prepracovanou terminológiou je však v korpuse národného jazyka stále nedostatok, a tak sa budeme postupne obracať aj na ďalších autorov zo Slovenskej akadémie vied, aby sa Slovenská terminologická databáza mohla stať skutočnou databázou obsahujúcou a istým spôsobom ustaľujúcou terminologické sústavy jednotlivých vedných odborov na Slovensku.

Oddelenie Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v Bratislave spolupracuje pri plnení konkrétnych vedeckých a výskumných úloh s pracoviskami na Slovensku i v zahraničí, každý druhý rok organizuje medzinárodnú konferenciu Slovko o počítačovom spracovaní prirodzených, predovšetkým slovanských jazykov (najbližšia bude 25. – 27. 10. 2007, http://korpus.juls.savba.sk/~slovko/2007/) a v plnom rozsahu napĺňa všetky možnosti na elektronizáciu jazykovedného výskumu na Slovensku a stabilné postavenie slovenčiny v medzinárodnom kontexte rozvoja informačných technológií. Pri hodnotení prvej fázy projektu to ocenili aj zahraniční posudzovatelia, keď Slovenský národný korpus zaradili medzi 5 – 8 špičkových jazykových korpusov na svete, aj vedenie SAV, keď tento kolektív

odmenilo v r. 2005 Cenou Slovenskej akadémie vied za budovanie infraštruktúry pre vedu.

MÁRIA ŠIMKOVÁ

(Autorka je pracovníčkou Jazykovedného ústavu Ľ. Štúra SAV)