Úvod

Zameranie práce

V posledných desaťročiach zaznamenala lingvistika búrlivý rozvoj, a to hlavne v oblasti tvorby a sprístupňovania veľkých, ľahko spracovateľných a prehľadávateľných zdrojov jazykových dát (predovšetkým textových korpusov). Vďaka nim sa dajú pomerne jednoducho a exaktne overovať alebo vyvracať hypotézy o jazyku z oblasti teoretickej lingvistiky. Zároveň ponúkajú mnohé ďalšie možnosti a nástroje, ktorými sa dajú jazykové prostriedky opisovať, analyzovať a spracovávať nad rámec klasických postupov.

Počítačové spracovanie prirodzeného jazyka sa stalo štandardným postupom pri riešení prakticky všetkých výskumných úloh, v ktorých je potrebné získavať informácie z väčšieho množstva textov, pričom nie je efektívne či reálne spracovať všetky vstupy ručne, a kde je k dispozícii dostatočné množstvo údajov na to, aby boli štatistické analýzy spoľahlivé. Dostupnosť rozsiahlych korpusov umožňuje použiť sofistikované matematické a štatistické metódy na tvorbu čoraz presnejších jazykových modelov a následnú spoľahlivejšiu analýzu rôznorodých aspektov jazyka. Okrem korektného použitia štatistických metód sa tak stáva dôležitým aj výber vhodného matematického modelu skúmaných javov.

Slovenčina je výrazne flektívny jazyk, preto sa v ňom gramatické kategórie vyjadrujú prevažne na konci tvarov slov príslušnými gramatickými morfémami. Súpis a analýza existujúcich zakončení slov spolu s informáciami o ich používaní v jazyku tak poskytuje užitočný základ pre prípadné štúdie predovšetkým v oblasti gramatiky, vo výskume slovenskej flexie, ako aj v derivatológii. V tomto kontexte sa javí ako nanajvýš potrebné spracovať aj zakončenia slov nielen v ich základnom tvare, ale aj vo všetkých tvaroch paradigmy. Existujúce retrográdne slovníky flektívnych jazykov spracovávajú takmer výlučne iba lemy (porovnaj *Literatúru*), a to, predpokladáme, už len čisto z pragmatických dôvodov (slovné tvary slovenčiny spracováva retrográdne Mistrík, 1985). Morfologická databáza Slovenského národného korpusu napríklad obsahuje viac ako milión unikátnych slovných tvarov, podobne aj korpus `prim-7.0-frk` obsahuje približne milión unikátnych slovných tvarov, ktoré nie sú hapax legomena. Hoci kvalitne spracovaná morfematická analýza základných slovných tvarov (cf. Ološtiak – Genči – Rešovská, 2015) dokáže do istej miery suplovať informácie o odvodených slovných tvaroch, takto spracované slovníky: 1) sú citlivé na prijaté princípy morfematickej analýzy, 2) vyžadujú si analýzu celého heslára, prípadne by si po zapracovaní štatistických údajov z korpusu vyžadovali kvalitnú a presnú automatickú morfematickú analýzu vstupných dát, 3) neumožňujú jednoduché vyhľadávanie podľa zakončení slov.

Predkladaný *Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu* (ďalej RS SNK) prístupným spôsobom zachytáva podstatnú časť tvarov slov s rozličnými zakončeniami, berúc na zreteľ rozsahové možnosti knižnej publikácie, ako aj prehľadnosť a možnosti efektívneho vyhľadávania v tomto špecifickom type slovníka. Slovník sa nesnaží nahrádzať či korigovať kodifikačné príručky, ale zachytáva jazyk v jeho prirodzenej podobe, v akej sa používa v štandardných písaných komunikátoch odborného, publicistického aj umeleckého štýlu. Dielo vzniklo na základe analýzy dát Slovenského národného korpusu, konkrétne špecializovaného korpusu `prim-7.0-frk`. Tento bol vytvorený ako referenčný korpus pre *Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu* (Garabík a kol., 2017; ďalej FSS SNK), pričom sa pri jeho tvorbe prihliadalo na možnosť využiť ho ako východiskový jednotný textový zdroj pre ďalšie lexikografické diela. Pri jeho príprave sa kládol dôraz na vyváženosť textov z hľadiska ich štýlovo-žánrového zaradenia so zámerom zachytiť relevantnú časť slovnej zásoby súčasnej slovenčiny (ide o texty z rokov 1991 až 2015). Bližšie informácie o korpuse `prim-7.0-frk` sa nachádzajú v úvode FSS SNK (s. 7 – 18).

Predpokladáme, že špecifická lexika typická pre texty rôzneho štýlového zaradenia neovplyvňuje vo výraznej miere frekvenciu zakončení slov, nakoľko jazyk je z gramatického hľadiska ustálený a viac-menej uniformný bez ohľadu na štýl komunikátov. Výber tohto korpusu ako referenčného zdroja pre RS SNK však okrem iného umožňuje konzistentné porovnanie s dátami z korpusov či informáciami zachytenými v ďalších lexikografických dielach, na ktoré tento slovník koncepčne voľne nadväzuje, a to na *Skloňovanie podstatných mien v slovenčine s korpusovými príkladmi* a *Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu*.

Štruktúra slovníka

V ďalšom texte budeme termínom slovo označovať konkrétny slovný tvar, pod počtom unikátnych slov budeme rozumieť počet rôznych slovných tvarov (kde každý je započítaný len raz) a pod počtom slova celkový počet výskytov daného slovného tvaru v korpuse. Koncový n-gram definujeme ako postupnosť n písmen, t. j. gramov (z gréckeho γραμμα – znak, písmeno) na konci slova. Slovník spracováva výlučne ortografickú rovinu jazyka, v ktorej sú základnými stavebnými prvkami slov grafémy. V súlade so zaužívanou tradíciou sa zložky *ch*, *dz*, *dž* definitoricky považujú za jedno písmeno.[1] Pre prehľadnosť uvádzame pred koncovým n-gramom znak - (spojovník). Pre slovo dĺžky n považujeme koncový n-gram za existujúci a rovný celému slovu. Takéto prípady sú v slovníku zámerne uvedené, pričom ponechávame znak spojovníka na začiatku n-gramu.

Slovník pozostáva z dvoch hlavných častí (*Početnosti koncových n-gramov*, *Koncové n-gramy podľa gramatických kategórií*) a z príloh. V hlavných častiach sa neuvádzajú skratky, citátové výrazy a neurčiteľné slovné druhy. Podobne sa neuvádzajú ani petrifikované časti viacslovných výrazov. V prvej časti sú v stromovej štruktúre usporiadané koncové n-gramy v lexikografickom abecednom usporiadaní. Pri každom n-grame uvádzame celkový počet výskytov daného n-gramu v korpuse a zoznam najčastejších slov končiacich na daný n-gram, usporiadaných podľa početnosti výskytu. Pri každom slove je uvedený počet jeho výskytov v korpuse. Za zoznamom slov je v zátvorke uvedená dvojica čísel: počet unikátnych zostávajúcich slov a celkový počet slov (t. j. počet výskytov daného n-gramu), ktoré sa do zoznamu nedostali. V prvej časti sa neuvádza žiadna dezambiguácia podľa slovných tried, a tak v prípade slov homonymných s neuvedenou slovnou triedou (napr. skratkami) je uvedený celkový počet tohto slova bez ohľadu na to, že niektoré výskyty by patrili do inej neuvádzanej slovnej triedy.

V druhej časti sú n-gramy rozdelené podľa svojej dĺžky (uni-, bi-, tri- a tetragramy), ďalej podľa zaradenia k slovným druhom a typu paradigmy; nasleduje rozčlenenie podľa relevantných gramatických kategórií rozlišovaných pri konkrétnych slovných druhoch. Unigramy a bigramy sú uvedené v plnom rozsahu, z trigramov a tetragramov je uvedený len výber najčastejších slov. Pri trigramoch uvádzame každý taký trigram, ktorého počet výskytov nie je menší než jedna šestnástina počtu výskytov najčastejšieho; pre tetragramy taký tetragram, ktorého počet výskytov nie je menší než jedna dvanástina počtu výskytov najčastejšieho. Pre každý n-gram je uvedený počet slov končiacich na daný n-gram, vrátane tých slovných druhov a gramatických kategórií, ktoré sa ďalej neuvádzajú. To znamená, že počet výskytov uvedených pri n-grame sa nemusí rovnať súčtu výskytov jednotlivých slov končiacich na daný n-gram.

Pre každú množinu gramatických kategórií je uvedený počet slov končiacich na daný n-gram v rámci uvedeného slovného druhu, paradigmy a hodnôt uvedených gramatických kategórií. Nasleduje graf znázorňujúci distribúciu počtu výskytov daného slova v závislosti od jeho rangu a parameter b Zipfovej-Mandelbrotovej distribúcie (pozri ďalej). Za dvojbodkou sa opäť nachádza zoznam najčastejších slov končiacich na daný n-gram s uvedeným počtom výskytov. Za zoznamom môže nasledovať (v zátvorkách) dvojica čísel označujúca počet unikátnych zostávajúcich slov a celkový počet slov, ktoré sa do zoznamu nedostali.

Abecedné triedenie n-gramov v slovníku sleduje princíp, podľa ktorého nasledujú písmená s diakritikou po písmenách bez diakritiky; zložka *ch* nasleduje po písmene *h*; zložky *dz* a *dž* nasledujú po písmene *d*. Takéto usporiadanie sa ukázalo ako prehľadnejšie a hlavne pri obrátenom usporiadaní (a tergo) vhodnejšie pri vyhľadávaní než zásady zaužívané pri tvorbe diel, zodpovedajúce už neplatnej norme STN 01 0181.

Gramatické členenie

Tradične sa kategorizácia slovných druhov v slovenčine (podobne ako v iných jazykoch) nezakladá výlučne na morfologických vlastnostiach slov. V mnohých prípadoch ju ovplyvňujú aj syntaktické, štylistické či sémantické charakteristiky a spôsoby použitia slov v kontexte. Často sa tak stretávame s nejednoznačnými hraničnými prípadmi, keď sa jednotlivé slová klasifikujú odlišne na základe uprednostňovania jedného z faktorov, čo sa prejavuje rozličnou interpretáciou a zaradením slov k rôznym slovným druhom v odbornej literatúre a v slovníkových dielach. Slovnodruhové zaradenie v dostupných opisoch gramatiky je preto z hľadiska gramatickej abstrakcie často nekonzistentné. Medzi problematické prípady pri formalizovanom spracovaní jazyka patria napríklad prípady, keď sa slovo používa „v platnosti iného slovného druhu“, prípadne je súčasťou zloženého výrazu.

V našom slovníku používame morfosyntaktický tagset Slovenského národného korpusu, ktorý sa vyznačuje vyššou mierou formalizmu a konzistentnejšou klasifikáciou slov v porovnaní s tradičnými opismi. Pre formálnu analýzu sa osvedčilo aj členenie slov podľa typu paradigmy, vďaka čomu vieme prehľadne zachytiť typické a menej typické slová v rámci slovného druhu (napríklad substantíva so substantívnou paradigmou sú neutrálne, ale substantíva s adjektívnou, zmiešanou či neúplnou paradigmou sú menej časté a z formálneho hľadiska príznakové). Typ paragidmy sa v slovníku uvádza v zátvorke za označením príslušného slovného druhu; v prípade, že je paradigma typická pre daný slovný druh, pre stručnosť ju neuvádzame (substantívnu paradigmu pri

[1] Okrem prípadov, keď sa tieto písmená nachádzajú na hranici dvoch morfém. Také slová sa ale v slovníku nevyskytli.

substantívach, adjektívnu pri adjektívach, zámennú pri zámenách atď.). V slovníku sú uvádzané slová a ich slovnodruhové zaradenie známe morfologickému analyzátoru.

Propriá sa vo všeobecnosti uvádzajú s veľkým začiatočným písmenom. Antroponymá, konkrétne krstné mená a priezviská, sa uvádzajú vždy s veľkým písmenom (s výnimkou takých, ktoré začínajú malým písmenom, prípadne obsahujú neobvyklú kombináciu veľkých a malých písmen). Toponymá a zoonymá sa uvádzajú s veľkým začiatočným písmenom, ak je dané slovo prevažne vlastným menom. V prípade, že sa používa aj ako apelatívum, slovo uvádzame s malým začiatočným písmenom. Toto pravidlo sa uplatňuje aj pri klasifikácii jednotlivých častí viacslovných pomenovaní, ak sa jednotlivé apelatíva vyskytujú v korpuse viac než zanedbateľne. V prípade slov fungujúcich ako apelatíva aj propriá sme odlíšenie ich zápisu reflektovali v prípade antroponým a jednoslovných názvov obcí. V prípade iných proprií zhodných s apelatívami sa zachováva rozdiel vo veľkosti začiatočného písmena, iba ak sa apelatívum a proprium líšia v gramatickom rode alebo čísle.

V ďalšej časti uvádzame stručný zoznam rozlišovaných slovných druhov a slovných tried, zoznam typov paradigiem a gramatických kategórií v nadväznosti na pravidlá kategorizácie uplatňované v Slovenskom národnom korpuse (Garabík – Šimková, 2012), pričom opisujeme aj niektoré osobitosti spracovania textových dát v tomto slovníku:

Zoznam slovných druhov	
Substantíva	Realizujú sa gramatické kategórie pádu (7 pádov, pričom vokatív zväčša, ale nie úplne, nahrádza tvar nominatívu), rodu (4 rody) a čísla (2 čísla). Pri substantívach s nesklonnou paradigmou zachytávame iba rod.
Adjektíva	Realizujú sa gramatické kategórie ako pri substantívach: rod (4), číslo (2), pád (6, tvar vokatívu nemá osobitné formy), stupne (2 – pozitív; a komparatív zlúčený so superlatívom). V prípade nesklonných adjektív gramatické kategórie nerozlišujeme, značíme len stupeň (pozitív).
Participoidy	Ako participoidy v slovníku vyčleňujeme slová, ktoré vznikli od príslušných tvarov slovies s použitím formantov využívaných na tvorbu particípií (odlišujú sa substantíva formálne zhodné s participoidmi). Táto trieda je formálnejšia a rozsiahlejšia ako tradične vyčleňovaná trieda particípií, zahŕňa teda viac motivátov ako klasicky uvádzané príčastia. Rozlišujú sa participoidy aktívne a pasívne. V slovenčine existuje aj činné príčastie minulé, ktoré sa používa veľmi zriedkavo, v slovníku kategóriu času pri činných participoidoch neznačíme.
Pronominá	Gramatické kategórie rozlišujeme v rámci konkrétnych typov paradigiem. Pri zámenách so zámennou paradigmou sa značí rod, číslo, pád. Aglutinované zámená (tvoriace s predložkou jedno slovo) sa tiež nachádzajú v tejto kategórii.
Numeráliá	Gramatické kategórie rozlišujeme v rámci konkrétnych typov paradigiem. Pri numeráliách s číslovkovou paradigmou sa realizuje gramatická kategória rodu a pádu.
Verbá	Určujeme slovesnú formu, pri všetkých slovesných formách sa realizuje kategória vidu. Pri l-ovom particípiu zachytávame kategóriu rodu a čísla, v pluráli len číslo. Pri tvaroch indikatívu, imperatívu a futúra sa značí osoba a číslo. Tvary *ahoj(te)*, *vitaj(te)*, *čau(te)* a podobné sa považujú za interjekcie, nie za verbá. Negatívne tvary slovies sú zlúčené s pozitívnymi, pričom prefix *ne-* sa neuvádza, okrem tvarov slovesa *nejsť*, v ktorých koreňová morféma (v indikatíve) má alternáciu -j-, a okrem tvarov *niet*, *nieto*, *neni*.
Adverbiá	Realizuje sa gramatická kategória gradácie. Označujeme ju rovnako ako pri adjektívach.
Prepozície	Rozlišujeme vokalizované a nevokalizované prepozície. Pre prehľadnosť sa ako prepozícia chápe aj slovo *à* (okrem prípadov, keď je súčasťou viacslovného spojenia alebo citátového výrazu).
Konjunkcie, partikuly, interjekcie a iné	Nerealizujú sa pri nich žiadne gramatické kategórie. Samostatne vyčleňujeme reflexívne *sa*, *si* a samostatnú kondicionálovú morfému *by* (ktorá je homonymná s konjunkciou *by* vo význame *aby*, vyskytujúcou sa ojedinele hlavne v umeleckých textoch, ale jej frekvencia je natoľko nízka, že ju nebolo potrebné dezambiguovať).
Ostatné slovné triedy	Patria sem skratky, citátové výrazy, rôzne neslovné elementy a neurčiteľné slovné druhy (napr. slová s preklepmi). V slovníku ich neuvádzame.

Zoznam typov paradigiem	
Substantívna	Realizuje sa pri substantívach, zámenách a číslovkách.
Adjektívna	Realizuje sa pri adjektívach, substantívach, zámenách a číslovkách.
Zmiešaná	Realizuje sa pri adjektívach, substantívach, zámenách a číslovkách. Je to typ paradigmy, pri ktorej sa jednoznačne nesledujú inflekčné vlastnosti jedného slovného druhu, ale v niektorých kategóriách sa preberajú vlastnosti iného slovného druhu. V slovníku priraďujeme zmiešanú paradigmu aj takým substantívam, ktoré sú nesklonné iba v jednom gramatickom čísle.
Neúplná	Realizuje sa pri niektorých substantívach, zámenách a číslovkách. Ide o slová, ktoré sa skloňujú len v časti paradigmy.
Nesklonná	Realizuje sa pri ohybných slovných druhoch. Slová majú vo všetkých pádoch a číslach nemenný tvar (zodpovedajúci leme).
Číslovková	Realizuje sa pri číslovkách.
Adverbiálna	Realizuje sa pri zámenách a číslovkách.
Pronominálna	Realizuje sa pri zámenách.

Zoznam gramatických kategórií	
Rod	Vzhľadom na odlišnosti v paradigme životných a neživotných maskulín sa tradične rozlišujú dva mužské gramatické rody: životný a neživotný. Celkovo sa vymedzujú 4 rody, a to mužský životný, mužský neživotný, ženský a stredný rod. V slovenčine (podľa kodifikačných príručiek) sa vyskytujú aj také neživotné maskulína, ktoré majú tvar v akuzatíve singuláru odlišný od nominatívu singuláru (*ducha*, *menovateľa*, *čitateľa* atď.). Tieto tvary pre prehľadnosť započítavame k životným pendantom.
Číslo	Rozlišujeme dve gramatické čísla: singulár a plurál. Tvary singulárií a plurálií tantum sú zachytené ako tvary singuláru alebo plurálu a skutočnosť, že druhé gramatické číslo pre ne neexistuje, nie je špeciálne vyznačená. Gramatická kategória čísla je vlastná slovesám, substantívam, adjektívam, zámenám a číslovkám.
Pád	Rozlišujeme 7 pádov: nominatív, genitív, datív, akuzatív, vokatív, lokál a inštrumentál. Tvar vokatívu sa vyčleňuje iba v prípade, že sa odlišuje od nominatívu.
Stupeň	Pri adjektívach, adverbiách a participoidoch rozlišujeme pozitív a spojenie komparatívu so superlatívom. Počty výskytov komparatívu a superlatívu sčítavame a v slovníku tvar komparatívu zastupuje obe formy.
Slovesná osoba	Rozlišujeme tri osoby: prvú, druhú a tretiu. Táto kategória je relevantná iba pri slovesných formách indikatívu, imperatívu (iba prvá a druhá osoba), l-ového particípia a futúra.
Slovesná forma	Pod týmto termínom rozumieme všeobecne morfologicky odlíšenú formu slovesa, zahŕňajúcu slovesný čas aj spôsob:
	infinitív — V slovenčine ho považujeme za základný tvar slovesa. Určujeme pri ňom len slovesný vid. Infinitív je tiež súčasťou zložených tvarov budúceho času, v tom prípade je budúci čas označený pri pomocnom slovese *byť*.
	indikatív — Pre stručnosť indikatív v slovníku neznačíme.
	imperatív — Slovesá v tvare imperatívu môžu byť len v 1. osobe singuláru alebo v 1. či 2. osobe plurálu.
	transgresív — Má iba jedinú formu.
	l-ové príčastie — Osobitná forma slovies, súčasť minulého času (s pomocným slovesom *byť*) a podmieňovacieho spôsobu (s morfémou *by* a pomocným slovesom *byť*).
	futúrum — Budúci čas, ktorý je vyjadrený morfologicky. Ide o budúci čas plnovýznamového slovesa *byť*, aj o pomocné sloveso *byť* ako súčasť zloženého budúceho času, ako aj budúci čas niektorých slovies pohybu, ktorý sa tvorí pomocou predpony *po-*.
Slovesný vid	Rozlišujeme dokonavý vid (perfektum) a nedokonavý vid (imperfektum). Slovesá, ktoré sa vyskytujú ako dokonavé aj nedokonavé, označujeme ako obojvidové. Sloveso *dať* uvádzame v záujme prehľadnosti ako obojvidové.
Slovesný rod	Rozlišujeme ho pri participoidoch, môže byť aktívny alebo pasívny.

Ručné korekcie

Slovenský jazyk sa, podobne ako ostatné slovanské jazyky, vyznačuje pomerne vysokým stupňom homonymie. Ide jednak o pravú homonymiu, t. j. výskyt identických slovných tvarov patriacich k rôznym lemám, ako aj o homonymiu tvarovú, t. j. identické tvary v rámci paradigmy. Určovanie gramatických kategórií sa vo všeobecnosti riadi princípmi prijatými v morfosyntaktickom značkovaní Slovenského národného korpusu (Garabík – Šimková, 2012). V slovníku bola využitá automatická lematizácia, morfologická analýza a dezambiguácia aplikovaná v Slovenskom národnom korpuse. Automatické spracovanie má, prirodzene, istú malú mieru chybovosti, v niektorých prípadoch nesprávne určuje gramatické kategórie alebo zaradenie proprií či apelatív. Hoci sú takéto prípady v absolútnych číslach málo časté, v slovníku sú zahrnuté aj slová s nízkymi výskytmi (na úrovni jednotiek) realizácií príslušných gramatických kategórií, pri ktorých je nesprávna morfologická anotácia výrazne viditeľná. Pristúpili sme preto k ručným korekciám výskytov niektorých javov.[2]

V úvodnej fáze prípravy slovníka sme vytvorili zoznam homonymných slovných tvarov, z ktorého sme v rámci konkrétnych koncových n-gramov vybrali vysoko frekventované tvary s výraznou homonymiou, ako aj tvary odlišujúce sa iba veľkosťou začiatočného písmena. Typicky sme vyselektovali vzorky (konkordancie) v rozsahu 20 výskytov. K týmto vybraným tvarom sme potom ručne priraďovali zodpovedajúce gramatické kategórie a podľa potreby zároveň upravovali veľkosť začiatočného písmena (prípadne viacerých písmen) slov. Súčasne sme pri korekciách označovali preklepy a citátové výrazy, o prislúchajúci pomer ktorých bol počet výskytov daného tvaru korigovaný.

Počet všetkých výskytov daného slovného tvaru (vrátane homonymných tvarov) vo všetkých možných slovenských textoch sa považuje za populáciu, v ktorej je potrebné odhadnúť podiel slovných tvarov so správne určenými gramatickými kategóriami. Podobne počet všetkých výskytov slovného tvaru s veľkým či malým začiatočným písmenom sa považuje za populáciu, v ktorej odhadujeme podiel slovných tvarov so správnou veľkosťou začiatočného písmena (po korekciách počtu výskytov voči výskytom na začiatku vety, prvé slová vo viacslovných názvoch a pod.).

Tento počet výskytov realizácií gramatických kategórií (alebo slov s danou veľkosťou začiatočného písmena) vo vzorke sa riadi hypergeometrickým rozdelením pravdepodobnosti, ktorého parametre odhadneme na základe vzorky. Vzhľadom na veľkosť populácie (t. j. všetkých možných slovenských textov) môžeme použiť aproximáciu hyperbolického rozdelenia pravdepodobnosti binomickým rozdelením. V slovníku sme použili Clopper-Pearsonov odhad intervalu spoľahlivosti (Pearson, 1895) pre binomické rozdelenie.

Pri počtoch výskytov, ktoré boli ručne korigované (t. j. pri danom tvare slova sa po kontrole ponechala správna gramatická informácia, prípadne sa nahradila správnou), za číselným údajom o počte výskytu slova nasleduje doplňujúca informácia: po znaku → (šípka doprava) nasleduje korigovaný počet, za ktorým sú uvedené hranice 90%-ného intervalu spoľahlivosti korigovaného počtu. Interval spoľahlivosti je uvedený ako vzdialenosť od očakávaného počtu v podobe dvoch čísel, ktoré treba pripočítať k očakávanej hodnote korigovaného počtu (spodná hranica je záporné číslo alebo nula, preto po pripočítaní k počtu bude spodná hranica intervalu menšia alebo rovná očakávanej hodnote). V prípade, že sa korigovaná hodnota rovná pôvodnej, je v záujme stručnosti uvedená iba raz (bez znaku →). Tento prípad môže nastať pri korekciách veľkosti prvých písmen slov, keď sa nezmení celkový počet výskytov slov končiacich na daný n-gram. V prípade, že sa vzdialenosť spodného okraja intervalu spoľahlivosti od očakávanej hodnoty rovná vzdialenosti horného okraja od očakávanej hodnoty, je namiesto oboch (v absolútnej hodnote rovnakých) vzdialeností uvedená len jedna, a to za znakom ± (plus-mínus).

ARF a homogenita výskytov

Samotný počet výskytov (frekvencia) nejakého javu (napríklad výskyt slova) v korpuse má vysokú výpovednú hodnotu, ale v niektorých prípadoch môže viesť k zavádzajúcim záverom. Ak sa napríklad slovo vyskytuje iba v istej obmedzenej časti korpusu (slovo nadužívané jedným autorom, prípadne objavujúce sa len v jednom type periodika a pod.), jeho frekvencia bude nadhodnotená v porovnaní s reálnym používaním slova v jazyku. V korpusovej lingvistike existuje a používa sa niekoľko rôznych spôsobov korigovania frekvencií tak, aby viac zodpovedali skutočnej výpovednej hodnote výskytov slova, prípadne sa používajú rôzne faktory disperzie slov v korpuse (Gries, 2008). V Slovenskom národnom korpuse je použitá priemerná redukovaná frekvencia (ďalej ARF[3]), ktorá je implementovaná v korpusovom manažéri (No)SketchEngine (Kilgariff et al., 2014).

V nasledujúcom značíme symbolom $arf(w)$ priemernú redukovanú frekvenciu výskytu w, symbolom $f(w)$ počet výskytov w v danom korpuse. V našom slovníku je w buď slovo, alebo kombinácia slova a realizácie gramatickej kategórie. Platí, že:

[2] Teoretické východiská sú podobné východiskám vo *Frekvenčnom slovníku slovenčiny na báze Slovenského národného korpusu*, hoci sa ich praktická realizácia vo výraznej miere odlišuje. Nasledujúci text je čitateľovi oboznámenému s FSS SNK z veľkej časti už známy.

[3] Average Reduced Frequency (Savický – Hlaváčová, 2002).

$$arf(w) = \frac{f(w)}{N} \sum_i \min\left(d_i, \frac{N}{f(w)}\right) \tag{1}$$

kde N je veľkosť korpusu, d_i je vzdialenosť medzi i-tým a i+1-ým výskytom slova w. ARF má niekoľko vlastností, ktoré ju predurčujú ako veľmi vhodnú na použitie v kvantitatívnej analýze lexiky: úzko súvisí so skutočnou frekvenciou; v prípade slova rovnomerne rozmiestneného v korpuse je s ňou totožná; v prípade slova koncentrovaného na tesne za sebou nasledujúcich pozíciách v korpuse sa zhora približuje k jednotke.

Ďalej si definujeme parameter homogenity (Garabík, 2017) ako:

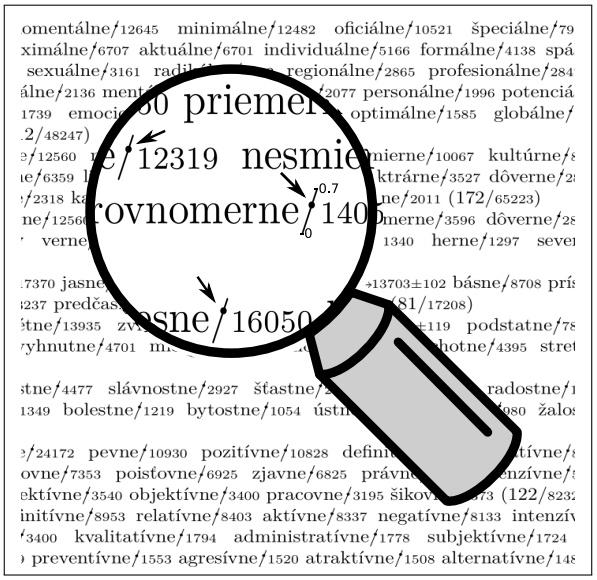
$$h(w) = \frac{arf(w) - 1}{f(w) - 1} \tag{2}$$

Do vzťahu pre výpočet homogenity vstupuje korekčný faktor -1, ktorý upravuje homogenitu tak, aby $h(w)$ bola definovaná len pre $f(w) \geq 2$, nakoľko pre $f(w) \in \{0, 1\}$ by bolo náročné intuitívne pochopiť zmysel tohto čísla, a upravuje rozsah pre slová s veľmi nízkym počtom výskytov. Korekčný faktor zároveň posunie hodnoty homogenity pre nízke (jednotkové) výskyty frekvencie k intuitívnejšiemu chápaniu výskytov rovnomerne rozložených v korpuse, ale iným významnejším spôsobom hodnotu neovplyvní. Homogenita má nasledujúce matematické vlastnosti:

- $0 < h(w) \leq 1$
- $h(w) = 1 \Leftrightarrow arf(w) = f(w)$, čiže homogenita sa tým viac blíži k hodnote 1, čím je slovo v korpuse rozmiestnené rovnomernejšie, a v hraničnom prípade je rovná číslu 1, ak je rozloženie výskytu slova úplne rovnomerné
- nízka hodnota $h(w)$ naznačuje, že výskyty slova sú niekedy navzájom bližšie k sebe, než je obvyklé.

Homogenita teda do istej miery odráža rovnomernosť rozloženia slova, empirické poznatky získané zo Slovenského národného korpusu ukazujú, že pre pomerne časté slová je zhora ohraničená hodnotou ≈ 0.66. Väčšiu homogenitu majú iba slová s pomerne malým počtom výskytov (t. j. jednotkami či desiatkami výskytov), kde je homogenita výsledkom štatistických fluktuácií, a teda pri takýchto nižších počtoch údaj o nej nie je spoľahlivým ukazovateľom.

Homogenita je v slovníku zobrazená vo forme indikátora (obr. 1) na oddeľovači počtu výskytov. V časti *Početnosti koncových n-gramov* ide o homogenitu výskytu slovných tvarov, v časti *Koncové n-gramy podľa gramatických kategórií* o homogenitu výskytu slovného tvaru v rámci danej gramatickej kategórie. Homogenita sa v oboch častiach slovníka uvádza pre nekorigovaný výskyt. Rozsah zobrazenej homogenity je vzhľadom na horeuvedené zistenia obmedzený na $0 \leq h \leq 0.7$. Homogenita väčšia ako 0.7 je zobrazená maximálnou hornou polohou indikátora, spodná pozícia indikátora zodpovedá homogenite $h = 0$, rozsah zobrazenia je lineárny. Indikátor homogenity nie je zobrazený, ak údaj o homogenite nie je k dispozícii (t. j. pri nedefinovanej homogenite ne(s)korigovaného počtu výskytov 0 alebo 1).



Obr. 1: Ukážka indikátorov homogenity (zvýraznených šípkami). Na ilustráciu je pri slove *rovnomerne* zobrazená aj stupnica homogenity. Hodnota homogenity slova *rovnomerne* je 0.473.

Zipfov a Zipfov-Mandelbrotov zákon

G. K. Zipf, ktorý sa považuje za priekopníka matematického výskumu distribúcie lexikálnych jednotiek v textoch, usudzoval, že v jazyku sa prejavuje princíp ekonómie, t. j. smeruje k úspornosti vyjadrovania (sa) – expedient sa snaží vyjadriť svoje myšlienky tak, aby vyvinul čo najmenšie možné úsilie (čo tiež zodpovedá všeobecnej tendencii ľudskej mysle inklinovať k lenivosti; por. Ferrero, 1894). Medzi jeho najznámejšie objavy patrí zákonitosť štatistického rozloženia, ktoré modeluje empirické pozorovanie ukazujúce, že poradie slova v zozname usporiadanom zostupne podľa početnosti slov (v danom texte, resp. korpuse) je nepriamo úmerné početnosti daného slova.

Ak označíme symbolom f frekvenciu (t. j. počet) výskytov slova, symbolom r jeho rang (poradie v zostupnom usporiadaní slov podľa frekvencie), potom podľa Zipfovho zákona približne platí

$$f = \frac{K}{r^b} \tag{3}$$

kde b je parameter distribúcie udávajúci rýchlosť poklesu, K multiplikatívna konštanta normalizujúca frekvenciu na veľkosť daného korpusu. Zovšeobecnením Zipfovho zákona je potom Zipfov-Mandelbrotov zákon, vyjadrený vzťahom

$$f = \frac{K}{(r+q)^b} \tag{4}$$

kde q je doplňujúci parameter, ktorý udáva istý posun rangu voči základnému poradiu v nemodifikovanom Zipfovom zákone.

Závislosť frekvencie a rangu sa zvykne znázorňovať v logaritmicko-logaritmickej mierke, v ktorej má vizuálne zobrazenie závislosti podobu (klesajúcej) priamky. V prípade nenulového parametra q bude tento pokles vizuálne „zakrivený“. V skutočnosti sa dá očakávať, že „naozajstná“ distribúcia nebude Zipfova ani Zipfova-Mandelbrotova, ale vznikne súčtom niekoľkých distribúcií, z ktorých každá je spôsobená iným fenoménom a týka sa inej množiny slov. Existuje niekoľko možných generalizácií Zipfovho zákona, ktoré zohľadňujú rôzny charakter poklesu početností v rôznych oblastiach podľa rangu alebo zdôvodňujú existenciu zákona vychádzajúc z teoretických poznatkov o početnostiach slov náhodne sa vyskytujúcich podľa istých rozdelení pravdepodobnosti (por. Sichel, 1975; Kornai, 1999; Montemurro, 2001; Piantadosi, 2014). Vzhľadom na to, že tieto generalizácie často skúmajú platnosť zákona v okrajových oblastiach nízkych výskytov alebo zavádzajú dodatočné voľné parametre rozdelení, je pre naše účely vzhľadom na často nízke výskyty koncových n-gramov s konkrétnymi gramatickými kategóriami vhodnejšie použiť jednoduchší Zipfov-Mandelbrotov zákon (4), ktorý aplikujeme na distribúciu počtu slov s daným koncovým n-gramom s konkrétnou gramatickou informáciou. Parameter b určuje rýchlosť poklesu frekvencie slov s ich rangom a má preto priamo výpovednú hodnotu, ktorú môžeme zhruba interpretovať ako inverzne korelujúcu s produktívnosťou danej prípony pre konkrétnu množinu gramatických kategórií.

Distribúcia slov zobrazená vo forme grafu (v spomínanej logaritmicko-logaritmickej mierke), ktorá sa nachádza pred zoznamom slov, môže slúžiť na rýchly náhľad tvaru poklesu početnosti slov podľa ich rangu. Hlavne sú v ňom viditeľné lokálne odchýlky od ideálneho mocninového poklesu alebo rozdiely vo fungovaní častých slov a málo častých slov. Na grafe je znázornený oddeľovač vo forme prerušenej zvislej čiary, ktorý oddeľuje pozíciu slov zachytených v zozname (naľavo od zvislej čiary) od zvyšných nezachytených slov (napravo). Ak sú všetky slová končiace na daný n-gram zachytené v zozname, oddeľovač nie je znázornený (v tom prípade by sa nachádzal v krajnej pravej časti grafu). Za grafom nasleduje hodnota parametra b získaná fitovaním distribúcie. Graf a parameter b sú uvedené len v tom prípade, že sa v danej gramatickej kategórii nachádzajú aspoň tri unikátne slová.

Prílohy

Zoznam koncových n-gramov zoradený podľa ich početnosti sa nachádza v prílohe *Zoznam koncových n-gramov podľa početnosti*.

V ďalšej prílohe *Najčastejšie počiatočné n-gramy* je zoznam najčastejších počiatočných n-gramov podľa ich početnosti, osobitne pre slovné tvary a osobitne pre lemy, samostatne sú uvedené analogické zoznamy pre plnovýznamové slovné druhy (pre prehľadnosť v nich neuvádzame zámená, oproti tomu tu uvádzame tvary slovesa *byť*). V zoznamoch počiatočných n-gramov (na rozdiel od predchádzajúcich častí slovníka) nie sú zlúčené tvary komparatívu a superlatívu a nie sú zjednotené ani pozitívne a negatívne tvary slovies.

V prílohe *Produktívnosť koncových n-gramov* sú tabuľky najproduktívnejších koncových n-gramov, pričom kvalitatívne produktívnosť koreluje inverzne s hodnotou parametra b distribúcie (4) – čím je parameter väčší, tým rýchlejšie klesá počet výskytov slov končiacich daným n-gramom s ich rangom. V tabuľkách je uvedený graf a hodnota parametra b distribúcie (4), počet unikátnych slov končiacich na daný n-gram a celkový počet daných n-gramov, slovný druh a gramatická kategória. Pre porovnanie, Mistrík (1985) definuje pri veľmi podobných

predpokladoch produktívnosť ako pomer celkového počtu slovných tvarov k počtu unikátnych slovných tvarov pre daný koncový n-gram (v našich tabuľkách recipročne 4. stĺpec).

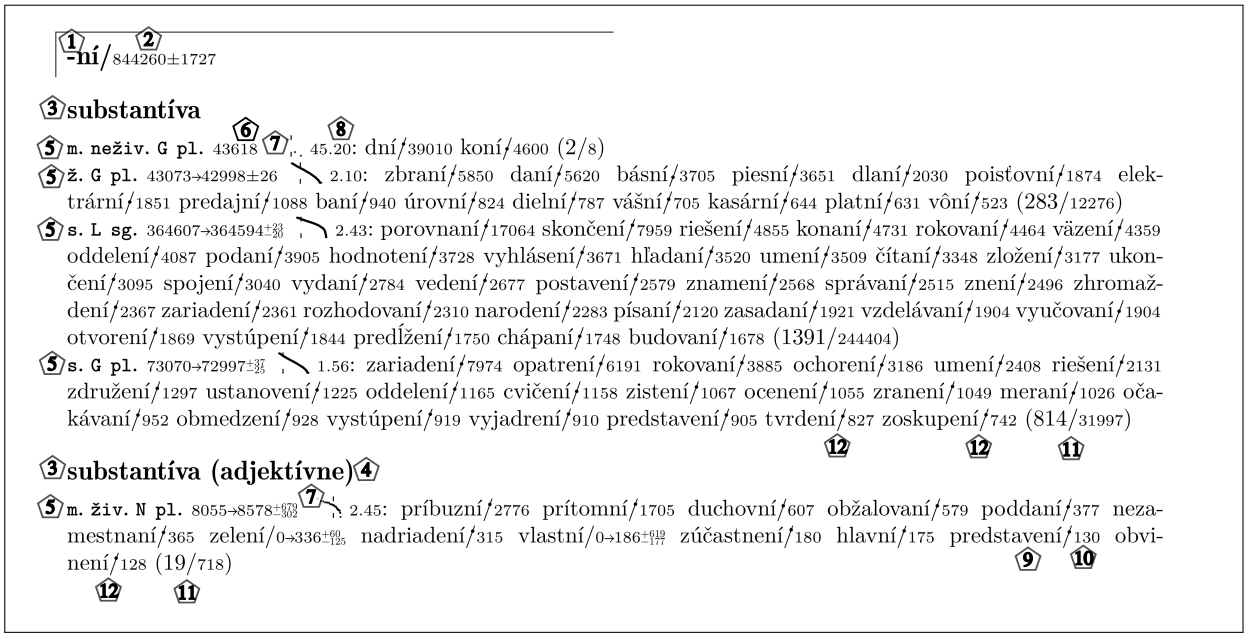
Tabuľky v poslednej prílohe *Prechody medzi n-gramami v rámci slov* zobrazujú pravdepodobnosti prechodov medzi jednotlivými n-gramami v rámci slova. Slová sú tak na ortografickej úrovni modelované Markovovým reťazcom n-tého rádu (Марковъ, 1908), kde stavom zodpovedajú unigramy (grafémy). Začiatok a koniec slova sa chápu ako samostatné stavy. Začiatok je označený symbolom ⍓→, koniec symbolom →⍓. Pravdepodobnosti prechodov sa uvádzajú v percentách. V strednom stĺpci je príslušný stav, v pravom stĺpci pravdepodobnosti prechodov k stavom v prográdnom smere (pravdepodobnosť, že za stavom v strede tabuľky nasleduje písmeno v pravej časti), v ľavom stĺpci pravdepodobnosti prechodov v retrográdnom smere (pravdepodobnosť, že pred stavom v strede tabuľky sa nachádza písmeno v ľavej časti). Umiestnený pod stavom v strede tabuľky je pomer výskytov daného stavu v celom korpuse ku všetkým stavom (bez citátových výrazov, ale so zahrnutím začiatkov a koncov všetkých ostatných slov do celkového počtu). Všetky počty sa vzťahujú na slová bez ručných korekcií, zahrnuté sú tu aj skratky, ale vynechané citátové výrazy. Napríklad riadok

67% ⍓→	p	r 35%

čítame tak, že zo všetkých výskytov písmena p je 67 % z nich predchádzané stavom začiatku slova (t. j. nachádzajú sa na začiatku slova) a za 35 % výskytov písmena p nasleduje písmeno r. Zobrazené sú len štyri najčastejšie prechody. Graficky sú prechody zobrazené na obr. 2 až obr. 7 (pozri prílohu). Pre bigramy a trigramy sme zoznamy skrátili a uviedli len najčastejšie z nich.

V Bratislave, október 2018 *Radovan Garabík, Agáta Karčová*

Štruktúra hesla



Legenda:

1. koncový n-gram (tu: bigram)
2. počet výskytov n-gramu
3. slovný druh
4. typ paradigmy
5. gramatické kategórie
6. počet výskytov slov končiacich na daný n-gram s danými gramatickými kategóriami
7. graf počtu výskytov slova v závislosti od jeho rangu
8. parameter b Zipfovej-Mandelbrotovej závislosti (jeho hodnota inverzne naznačuje produktívnosť daného zakončenia slova)
9. slovo končiace na daný n-gram s danými gramatickými kategóriami
10. počet výskytov slova
11. počet unikátnych slovných tvarov nezahrnutých do zoznamu slov / celkový počet slov nezahrnutých do zoznamu slov
12. ukazovateľ homogenity výskytov slova