GRZEGORCZYKOWA, Renata – PUZYNINA, Jadwiga: Problemy ogólne słowotwórstwa. W: Gramatyka współczesnego języka polskiego. t. 2. Morfologia. Red. R. Grzegorczykowa, R. Laskowski, H. Wróbel. Warszawa: PWN 1998, s. 361 – 388.

KAPROŃ-CHARZYŃSKA, Iwona: Derywacja ujemna we współczesnym języku polskim. Rzeczowniki i przymiotniki. Toruń: TOP KURIER 2005. 221 s.

KAPROŃ-CHARZYŃSKA, Iwona: Kilka uwag o derywatach tautologicznych. W: Studia nad słownictwem dawnym i współczesnym języków słowiańskich. Toruń: w druku.

KURDYŁA, Tomasz: Luksus słowotwórczy, czyli o tzw. derywatach tautologicznych i funkcjach tworzących je formantów. W: Język Polski, 2002, rocznik LXXXII, z. 3, s. 178 – 187.

KURYŁOWICZ, Jan: Językoznawstwo strukturalne. Warszawa 1979.

NAGÓRKO, Alicja: Zarys gramatyki polskiej (ze słowotwórstwem), Warszawa: PWN 1998. 331 s.

NAGÓRKO, Alicja: Zarys gramatyki polskiej. Warszawa: PWN 2003a. 331 s.

NAGÓRKO, Alicja: Tendencje w sferze pragmatyczno-stylistycznej. Różnicowanie i unifikacja środków słowotwórczych w służbie pragmatyki. W: Komparacja współczesnych języków słowiańskich. Słowotwórstwo/Nominacja. Red. I. Ohnheiser. Opole: Uniwersytet Opolski 2003b, s. 189 – 97 i 217 – 33.

SKARŻYŃSKI, Mirosław: Powstanie i rozwój polskiego słowotwórstwa opisowego. Kraków: TAiWPN Universitas 1999. 215 s.

ZEMSKAJA, Elena: Słowoobrazowanije kak diejatielnost'. Moskwa: Nauka 1992, 221 s.

Príprava a uskutočňovanie projektu morfologického analyzátora

Agáta Karčová

Slovenský národný korpus, Jazykovedný ústav Ľudovíta Štúra, SAV, Bratislava

Morfologický analyzátor je počítačový program, ktorý v sebe zahŕňa všetky tvary slovenských lexém vrátane ich tagov (súboru morfologických značiek). Projekt morfologického analyzátora sa začal uskutočňovať v oddelení Slovenského národného korpusu JÚĽŠ SAV v Bratislave v auguste 2005. Po vytvorení počítačového programu, v ktorom je možné uchovávať lexémy roztriedené podľa slovných druhov, vytvárať a priraďovať lexémam nové vzory, sa začala práca na dopĺňaní slovníka pre tento morfologický analyzátor.

Základným zdrojom materiálu na spracovanie bol *Krátky slovník slovenského jazyka* (2003; ďalej KSSJ), ktorý sme mali k dispozícii v elektronickej podobe. Lexémy z KSSJ sa roztriedili podľa ich príslušnosti k jednotlivým slovným druhom. Ku každej lexéme bol následne priradený vzor v súlade s koncepciou, ktorá je opísaná v publikácii *Model morfologickej databázy slovenčiny* (Benko – Hašanová – Kostolanský, 2004). Počet vzorov v dátach bol oveľa vyšší opro-

ti vzorom zaužívaným v tradičnej morfológii (233 nových vzorov oproti dvanástim základným vzorom pre substantíva), napriek tomu nebol pre naše zámery a ciele dostatočný. Pri každej lexéme sme preto museli overovať opodstatnenosť vzoru a súbor pôvodných vzorov modifikovať a dopĺňať.

Vo východiskovej koncepcii slovníka pre morfologický analyzátor sa uprednostnil formálny prístup, tak ako pri súbore pravidiel na ručnú anotáciu, čo však so sebou prinášalo výhody aj nevýhody. K výhodám nesporne patrí samotná možnosť vytvoriť funkčný a prehľadný elektronický slovník. Uplatňovanie formálneho prístupu urýchľuje realizáciu slovníka pre morfologický analyzátor, umožňuje prehľadnosť, presnosť a urýchlenie vyhľadávania všetkých tvarov slovenských lexém. Jednotlivé lexémy sú v tomto slovníku zachytené ako izolované ucelené jednoslovné jednotky, preto pomocou tohto programu nie je možné naraz zachytiť a vyhľadávať analytické tvary slov (napr. *bol by som robil, budeme sa stretávať*). Program však umožňuje nájsť každú časť týchto tvarov ako samostatnú jednotku.

Homonymia je v slovníku zachytená len v tých prípadoch, keď sa prejavuje aj formálne. Ako príklad uvádzame slovo *behúň*, ktoré má rôzne významy: ako životné substantívum 1. rýchly bežec, 2. rýchly kôň alebo iné zviera; ako neživotné substantívum 1. pohyblivá súčasť prístroja, 2. dlhý úzky pokrovec (podľa KSSJ). Substantívum *behúň* má tri rôzne ucelené paradigmy, v ktorých sa gramatické relačné morfémy v niektorých pádoch líšia, rovnako jeden znak v tagu je rozdielny (pri substantívach mužského rodu rozlišujeme životnosť a neživotnosť). Tieto odlišnosti sú v slovníku pre morfologický analyzátor zachytené tak, že slovu *behúň* sú priradené tri rôzne vzory: vzor *kráľ* pre 1. význam životného substantíva, vzor *jeleň* pre 2. význam životného substantíva a vzor *grúň* pre ostatné významy. Slovo *hlava* má takisto viacero rôznych významov, tie sa však v tvaroch paradigmy nevyznačujú nijakými rozdielmi, preto sme tomuto slovu, prihliadajúc primárne na formu, pridelili len jeden vzor *strava.*

Using paradigm: grúň	Using paradigm: jeleň	Using paradigm: kráľ	Using paradigm: strava
SSip1: behúne	SSip1: behúne	SSmp1: behúni	SSfp1: hlavy
SSip2: behúňov	SSip2: behúňov	SSmp2: behúňov	SSfp2: hláv
SSip3: behúňom	SSip3: behúňom	SSmp3: behúňom	SSfp3: hlavám
SSip4: behúne	SSip4: behúne	SSmp4: behúňov	SSfp4: hlavy
SSip5: behúne	SSip5: behúne	SSmp5: behúni	SSfp5: hlavy
SSip6: behúňoch	SSip6: behúňoch	SSmp6: behúňoch	SSfp6: hlavách
SSip7: behúňmi	SSip7: behúňmi	SSmp7: behúňmi	SSfp7: hlavami
SSis1: behúň	SSms1: behúň	SSms1: behúň	SSfs1: hlava
SSis2: behúňa	SSms2: behúňa	SSms2: behúňa	SSfs2: hlavy
SSis3: behúňu	SSms3: behúňovi	SSms3: behúňovi	SSfs3: hlave
SSis4: behúň	SSms4: behúňa	SSms4: behúňa	SSfs4: hlavu
SSis5: behúň	SSms5: behúň	SSms5: behúň	SSfs5: hlava
SSis6: behúni	SSms6: behúňovi	SSms6: behúňovi	SSfs6: hlave
SSis7: behúňom	SSms7: behúňom	SSms7: behúňom	SSfs7: hlavou

Obrázok 1: Ukážka paradigiem slova *behúň* v rôznych významoch a paradigmy slova *hlava.* Jednotlivé tvary sú uvedené tagom, ktorý označuje substantívum so substantívnou paradigmou mužského rodu – životné (i) alebo neživotné (m), ženského rodu (f) v singulári alebo pluráli v príslušnom páde.

Časť textov Slovenského národného korpusu už je ručne morfologicky označkovaná a keďže analyzátor bude použitý hlavne na automatizované označkovanie všetkých textov Slovenského národného korpusu, bolo potrebné pri dopĺňaní slovníka a tvorbe vzorov prihliadať na zaužívaný súbor pravidiel a značiek, tzv. tagset, ktorého aktuálna verzia *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu* je k dispozícii na URL: http://korpus.juls.savba.sk.

Rozdiely oproti zaužívanému triedeniu slov sa prejavujú v tom, že formálne particípiá sú vyčlenené ako samostatná trieda slov, za príčastia považujeme aj adjektivizované príčastia typu *vynikajúci*, *napísaný, vzdialený, krytý*.

Do systému pádov v morfologickom značkovaní menných slovných druhov bol zaradený aj vokatív. Analýza väčšieho množstva slovenských textov ukazuje, že vokatív nevymizol úplne a pri našej snahe o úplnosť a presnosť bolo opodstatnené jeho opätovné zaradenie do systému pádov. Jednou zo špecifických informácií v tagoch pre nominálne slovné druhy je tiež uvádzanie informácie o type paradigmy na druhom mieste v tagu bezprostredne po určení slovného druhu. Napríklad pri substantívach začiatok tagu môže vyzerať takto: *SS* – substantívum so substantívnou paradigmou (*mama*), *SA* – substantívum s adjektívnou paradigmou (*vedúci*), *SF* – substantívum so zmiešanou paradigmou (*kuli*), *SU* – substantívum s neúplnou paradigmou (*kupé*).

V priebehu vytvárania a dopĺňania slovníka pre morfologický analyzátor sa vyskytli mnohé problémy, ktoré vyplývajú zo zložitosti jazykového systému slovenského jazyka a mnohých nesystémových javov, ktoré sa v gramatike vyskytujú, resp. sú spôsobené dynamikou jazyka či dynamikou jeho gramatickej zložky. Ďalšie problémy boli s mnohými hraničnými a okrajovými javmi, ktorých riešenie je nejednoznačné a vo viacerých relevantných prameňoch sú interpretované rôzne, alebo naopak nie sú doposiaľ hĺbkovo spracované a nevenuje sa im dostatočná pozornosť.

Bolo potrebné riešiť aj to, či a do akej miery sa budú zohľadňovať zmeny v paradigmách vyplývajúce z čiastočnej zmeny pravidla o rytmickom krátení (Pravidlá slovenského pravopisu, 1991). Vzhľadom na to, že náš slovník vychádza z tvarov uvedených v kodifikačnej príručke *Pravidlá slovenského pravopisu* (2000), zmeny v rytmickom krátení sme v plnej miere rešpektovali a zachytávali. Z nášho uhla pohľadu a vzhľadom na praktické zameranie morfologického slovníka však bolo dôležité spracovať aj také tvary slov, ktoré sa používajú v písaných textoch súčasného obdobia a v istom čase boli z hľadiska normy správne. Preto sme okrem dnes platného tvaru zaradili do zoznamu slov aj všetky ostatné tvary tak, ako sa reálne vyskytujú v textoch Slovenského národného korpusu, mapujúcich slovnú zásobu súčasnej slovenčiny, t. j. od r. 1955 (*bábkár*/*bábkar*, *mliekáreň*/*mliekareň*, *lúkárstvo*/*lúkarstvo*, *lúpáreň*/*lúpareň*, *žiháreň*/*žihareň* a pod.).

Dvojrodovosť niektorých substantív (*skalisko, knieža, show* a iné) sme zachytili takým spôsobom, že sme každé zo slov zaradili k obom príslušným

rodom s vlastnými paradigmami a tagmi. Zaujímavé bolo z tohto hľadiska slovo *knieža*, ktoré je vnímané ako dvojrodové iba v nominatíve singuláru, preto sme preň vytvorili osobitný vzor mužského rodu, ktorý má len tento jeden tvar. Slovo knieža chápané ako substantívum stredného rodu má paradigmu úplnú.

Osobitným spôsobom bolo potrebné postupovať pri dopĺňaní paradigiem slov cudzieho pôvodu *capriccio*, *bolero*, *addagio* a iných. Podľa Morfológie slovenského jazyka (1966) má byť v G plurálu tvar *capriccií*, *boler*, tomu však nezodpovedá príslušný gramatický aparát v KSSJ a iných slovníkoch, kde sú uvedené tvary *capríč*, *bolier*. Rôzne interpretované sú aj tvary L plurálu *capricce* verzus *capricciu* a iné. Na vyriešenie tohto problému nám výrazne napomohol Slovenský národný korpus a rôzne internetové portály, v ktorých sme overovali výskyty jednotlivých tvarov.

Pri niektorých substantívach ženského a stredného rodu bolo náročné zistiť formu G plurálu, napr. pri substantívach *mäta, mokka, päta, honba, faloš, krv, hniloba, borba, fiasko* a pod. Pri slove *mokka* sa vyskytol navyše problém so zápisom slova v tomto tvare (ponechanie dvoch písmen *k* na konci tvaru slova alebo len jedného). Po konzultáciách s pracovníkmi Oddelenia jazykovej kultúry a terminológie JÚĽŠ SAV sme do príslušných paradigiem doplnili tvary *miat*, *mokk*, *piat*, *honieb*, *falší*, *krví*, *hnilôb*, *borb*, *fiask.*

Náročné bolo dopĺňanie paradigiem málo frekventovaných slov, ktoré sú súčasťou frazeologizmov, ich význam bez kontextu je už zastretý a bežnému používateľovi jazyka neznámy. Bolo potrebné zvážiť, či je nevyhnutné vytvárať plnú paradigmu pre slová typu *forota* (*mať vo forote* znamená mať v zásobe), *lub* (*mať za lubom* znamená zamýšľať niečo) a pod. Vzhľadom na to, že vytvorenie plných paradigiem týchto slov bolo bezproblémové a rôzne tvary týchto slov sa môžu vyskytovať v písaných textoch, vytvorili sme pre ne úplné paradigmy.

Pri tvorbe slovníka sme uprednostňovali formálny prístup, ale s tesným prepojením a neustálym prihliadaním na významovú zložku jazyka. Dôležité bolo napríklad uvedomenie si významu slov *fras, škriatok, duch, šarkan, čert, škrabák,* ktoré môžu byť použité vo význame životného i neživotného substantíva, čomu zodpovedajú aj príslušné rozdiely v ich paradigmách.

V priebehu práce sme odhalili aj niektoré nedôslednosti v doteraz známom spracovaní jednotlivých lexém, napr. pri slove *kôrovie* chýba relevantný údaj, že ide o hromadné podstatné meno. Pri pomnožnom substantíve *gate* je uvedený aj jeho ekvivalent v jednotnom čísle *gaťa,* definovaný ako *„časť gatí zakrývajúca jednu nohu“*, pri pomnožnom substantíve *nohavice* však takýto ekvivalent uvedený nie je, a pritom sa tento tvar vo význame *časť nohavíc zakrývajúca jednu nohu* používa a reálne vyskytuje (uvádzame niektoré výsledky vyhľadávania zo Slovenského národného korpusu: *jedna nohavica je nasiaknutá krvou; mokrá nohavica ho chladila; pravá nohavica; vzadu sa má nohavica začínať tam* atď.).

Rozsiahlym problémom dotýkajúcim sa spracovania takmer všetkých substantív mužského rodu zakončených na dve a viac spoluhlások bolo určenie správneho tvaru I. plurálu (zakončenie *-mi, -ami*) pri okruhoch vzorov *chlap*, *dub* a *stroj*. Táto problematika nie je dostatočne teoreticky rozpracovaná a ak sa v niektorých čiastkových štúdiách tento jav aspoň čiastkovo rieši, jednotliví autori sa vo formulovaní pravidiel na tvorbu správneho tvaru I plurálu podstatne líšia. Pravidlá sú nejednoznačné až vágne a v mnohých prípadoch nezohľadňujú reálne výskyty, ale za správnu považujú len jednu koncovku bez pripustenia dvojtvarov. Tieto pravidlá sú pritom formulované tak zložito, že sa v nich bežný používateľ nezorientuje. Riešením by bolo uvádzať príslušné prípony v novej verzii slovníka. Tvary I plurálu v KSSJ nie sú zachytené a v Pravidlách slovenského pravopisu (2000) je správny tvar uvedený iba pri niektorých z nich. Pri tvorbe morfologického slovníka sme neustále overovali reálny stav používania v súčasných textoch. Zistili sme, že prípona *-mi* je pri substantívach mužského rodu primárna a natoľko silná, že preniká aj do paradigiem substantív patriacich do deklinačného okruhu so vzorovým slovom *hrdina* (napr. *jezuitami, banditami* a pod.), tu je však tento tvar vnímaný ako príznakový. V okruhu vzorov *chlap, dub* a *stroj* je situácia oveľa zložitejšia. Pravidlo, že substantíva mužského rodu končiace na *-nt, -nd, -rd, -rt, -kt* majú mať v I singuláru príponu *-mi* (Navrátil, 2002), neodráža reálny stav, ako to zreteľne vyplýva z nasledujúcich príkladov v tabuľke. Náš postup pri zisťovaní prípon bol časovo náročný – overovali sme výskyty na portáloch Morfeo, Google a v Slovenskom národnom korpuse (http://korpus.juls.savba.sk.) a podľa počtu výskytov a iných relevantných kritérií sme zaraďovali jednotlivé slová k jednému alebo dvom vzorom.

tvary slov	**Google**	**prim-2.0**	**Morfeo**
absolventmi/absolventami	826/927	6/111	859/994
abstinentmi/abstinentami	14/67	0/4	7/14
abstraktmi/abstraktami	93/51	0/4	57/20
adeptmi/adeptami	187/286	11/78	89/219
agentmi/agentami	919/991	12/154	335/705
architektmi/architektami	179/589	10/58	148/377
asistentmi/asistentami	176/761	12/146	100/491
atramentmi/atramentami	285/315	0/0	1 012/193
bastardmi/bastardami	19/10	0/2	18/11
bleskmi/bleskami	731/23	54/11	353/11
boxmi/boxami	411/161	15/13	204/81
brejkmi/brejkami	114/89	42/47	74/64
brestmi/brestami	25/3	3/2	16/1
briliantmi/briliantami	37/166	0/32	23/111

Tabuľka 1: Počet výskytov jednotlivých tvarov substantív mužského rodu v I. plurálu (SNK je Slovenský národný korpus).

Zaraďovanie prídavných mien do vzorov bolo menej náročné, vyskytli sa tu však iné špecifické problémy, ako napríklad vyčlenenie stupňovateľných a nestupňovateľných prídavných mien. Zaujímavé boli aj výsledky vyhľadávania tvarov komparatívu adjektív, pri ktorých sme predpokladali dva variantné spôsoby tvorenia – pomocou prípony *-ší* alebo *-ejší.* Na základe nájdených výskytov sme každému z týchto adjektív priradili jeden alebo dva vzory. Niektoré tvary sa však nevyskytovali vôbec, preto sme sa aj po zohľadnení iných skutočností rozhodovali v prospech priradenia do toho vzoru, ktorý bol reálnejší, prípadne sme mu priradili dva vzory.

Slovo	**Google**	**prim-2.0**	**Vzor(y)**
bezduchší/bezduchejší	0/4	0/0	ostrý
prudší/prudkejší	757/9	99/5	hladký, ostrý
bližší/blízkejší	44 100/6	12 528/0	blízky
jednoduchší/jednoduchejší	10 300/6	1 980/0	bohatý
Hluchší/hluchejší	7/20	13/2	ostrý, bohatý
lichší/lichejší	0/0	0/0	ostrý, bohatý
Plachší/plachejší	23/12	0/4	ostrý, bohatý
Plochší/plochejší	44/68	8/9	ostrý, bohatý
suchší/suchejší	606/1	217/0	bohatý
svetloplachší/svetloplachejší	0/0	0/0	ostrý, bohatý
tichší/tichejší	1 320/0	598/0	bohatý
Vetchší/vetchejší	0/0	0/0	ostrý
vlhší/vlhkejší	101/66	30/28	ostrý, krotký
krotší/krotkejší	20/19	2/5	ostrý, krotký

Tabuľka 2: Výskyty jednotlivých tvarov adjektív v komparatíve a priradené vzory.

V súčasnosti sú už pre morfologický analyzátor spracované všetky slová patriace do menných slovných druhov vrátane osobitne vyčlenených participií. Do konca marca 2006 plánujeme spracovať aj slovesá a všetky neohybné slovné druhy, čím sa pokryje celá slovná zásoba zachytená v KSSJ.

Morfologický analyzátor bude primárne slúžiť na automatickú lematizáciu, morfologickú anotáciu a dezambiguáciu textov Slovenského národného korpusu. Takýmto spôsobom bude k dispozícii presnejší materiál na ďalší vedecký výskum slovenského jazyka. Morfologický slovník môže slúžiť aj ako východiskový podklad na prehodnotenie sústavy vzorov ohybných slovných druhov, predovšetkým substantív a pomôcť pri vytvorení novej koncepcie zatrieďovania slov do vzorov. Analyzátor a generátor tvarov môže byť prínosným aj v širšom rámci, napríklad ako základ pre databázové systémy určené na vyhľadávanie tvarov slovenských lexém v rôznych kultúrnych a vzdelávacích inštitúciách, knižničných systémoch a pod. V neposlednom rade môže byť využitý bežnými používateľmi jazyka a slúžiť ako cenná pomôcka na skvalitnenie a efektivizáciu výučby slovenčiny ako cudzieho jazyka.

Slovník pre morfologický analyzátor v súčasnosti spracúva len malú časť slovnej zásoby – jej najfrekventovanejšie slová a tvary z jadra slovnej zásoby. Po spracovaní lexém z KSSJ plánujeme elektronický slovník značne rozšíriť nielen o ďalšie slová domáceho pôvodu, ale aj o internacionalizmy, termíny z rozličných vedných oblastí, cudzie slová, neologizmy, vlastné mená, geografické názvy a iné. Zvyšovanie kvantity slovníka pre morfologický analyzátor opäť rozšíri možnosti jeho využitia.

Literatúra

BENKO, Vladimír – HAŠANOVÁ, Jana – KOSTOLANSKÝ, Eduard: Model morfologickej databázy slovenčiny. Počítačové spracovanie jazyka. Trnava: Univerzita sv. Cyrila a Metoda 2004. 190 s.

DVONČ, Ladislav: Dynamika slovenskej morfológie. Bratislava: Veda 1984. 124 s.

DVONČ, Ladislav: a kol.: Morfológia slovenského jazyka. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1966. 886 s.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. http://korpus.juls.savba.sk/publications/

Krátky slovník slovenského jazyka. 4., doplnené a upravené vydanie. Bratislava: Veda 2003. 986 s.

MIKO, František: Rod, číslo a pád podstatných mien. Bratislava: Vydavateľstvo SAV 1962. 256 s.

NAVRÁTIL, Ladislav: Inštrumentál množného čísla podstatných mien. In: Kultúra slova, 2002, roč. 36, č. 3, s. 139 – 144.

NAVRÁTIL, Ladislav: Nominálne slovné druhy. Nitra: Enigma 2002. 148 s.

ORAVEC, Ján – BAJZÍKOVÁ, Eugénia – FURDÍK, Juraj: Súčasný slovenský spisovný jazyk – morfológia. Bratislava: SPN 1988. 232 s.

Pravidlá slovenského pravopisu. 3., upravené a doplnené vydanie. Bratislava: Veda 2000. 592 s.

SOKOLOVÁ, Miloslava: Kapitolky zo slovenskej morfológie. Prešov: Slovacontact 1995. 180 s.

STANISLAV, Ján: Dejiny slovenského jazyka II. Tvaroslovie. Bratislava: Vydavateľstvo SAV 1985. 744 s.

Slovenský národný korpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2005. Dostupný z WWW: http://korpus.juls.savba.sk