DISKUSIE A ROZHĽADY

LEMATIZÁCIA, MORFOLOGICKÁ ANOTÁCIA A DEZAMBIGUÁCIA SLOVENSKÉHO TEXTU – WEBOVÉ ROZHRANIE

Radovan Garabík – Kristína Bobeková
Jazykovedný ústav Ľ. Štúra SAV
Bratislava

GARABÍK, Radovan – BOBEKOVÁ, Kristína: Lemmatization, Morphological Annotation and Disambiguation of the Slovak Text – Web Interface. Slovak Language, 2021, Vol. 86, No. 1, pp. 104 – 109.

Abstract: Lemmatization, morphological (or morphosyntactic) annotation (MSD) and disambiguation is a basic and indispensable step in Natural Language Processing of languages with a moderate level of inflection. We present a web interface demonstrating the de facto default lemmatization and MSD for Slovak, as used in major Slovak corpora (with several enhancements yet to be applied in the corpora). The interface can be used chiefly for presentation or pedagogical purposes, with the morphological tags expanded and explained using plain language in several languages, including two different terminological registers of Slovak (professional linguistic or a "common" one).

Key words: lemmatization, MSD, POS tagging, Slovak, web interface, morphological annotation, NLP

Jazykové technológie sú informačné technológie, ktoré sa zameriavajú na prácu s ľudským jazykom. Ide o oblasť úzko spojenú s počítačovým spracovaním prirodzeného jazyka (NLP), ktoré je vedeckou oblasťou na pomedzí lingvistiky, informatiky a v dnešnej dobe sa často považuje za oblasť umelej inteligencie.

Jazykovedný ústav Ľ. Štúra SAV patrí k popredným vedeckovýskumným inštitúciám na Slovensku zaoberajúcim sa počítačovým spracovaním slovenčiny a jazykovými technológiami. V rámci výskumu a tvorby nástrojov na počítačové spracovanie slovenčiny ústav poskytuje verejnosti niekoľko nástrojov demonštrujúcich rôzne aspekty spracovania jazyka.

Základom spracovania jazykov so strednou úrovňou flexie (ku ktorým patria vo všeobecnosti slovanské jazyky) je lematizácia a morfologická anotácia (značkovanie).

Pod lematizáciou sa rozumie určenie základného tvaru slova (lemy) patriacemu k danému slovu, resp. formálne ako priradenie istého unikátneho reťazca znakov identifikujúceho lexému. Lema (v korpusovom ponímaní) nemusí byť nevyhnutne

základným tvarom slova z klasickej lingvistickej analýzy, ani jedným z tvarov lexémy, hoci v popisovanej analýze slovenčiny takmer vždy je[1].

Morfologická anotácia znamená priradenie konkrétnych hodnôt slovných druhov a gramatických kategórií k jednotlivým slovám. Často sa stretávame aj s pojmom značkovanie POS (určenie slovného druhu alebo slovnodruhová anotácia; z anglického Part of speech (tagging)); kým v jazykoch s menším stupňom flexie (napr. čínština alebo angličtina) ide často o postačujúcu a často s lematizáciou aj jedinú vykonanú analýzu, pri slovenčine je samotné určenie slovného druhu len prvým krokom a je nedostatočné pre komplexné spracovanie jazyka.

V prípade slovanských jazykov sa pomerne univerzálne spája lematizácia s morfologickou anotáciou, keďže model *lema/morfologická značka/slovný tvar* veľmi dobre vystihuje morfologickú podstatu flektívnych jazykov[2], tento proces sa niekedy zastrešuje aj názvom morfologická analýza, hoci striktne povedané, výsledkom morfologickej analýzy nemusí byť lema, aj keď v prípade našich jazykov analýza dostatok informácie na priradenie lemy k slovnému tvaru poskytuje.

V skutočnosti lematizácia a morfologická anotácia môže poskytovať viacero možností; ako príklad uveďme slovenské slovo *zdraví*, ktoré môže byť:

1. lokál singuláru podstatného mena *zdravie*;
2. genitív plurálu podstatného mena *zdravie*;
3. nominatív plurálu prídavného mena *zdravý*;
4. nominatív plurálu podstatného mena[3] *zdravý*;
5. tretia osoba indikatívu slovesa *zdraviť*.

V takom prípade sa niekedy osobitne vyčleňuje aj ďalší stupeň spracovania, a to morfologická dezambiguácia („zjednoznačnenie“), kde sa z niekoľkých možností vyberá „tá správna“, najčastejšie použitím štatistických alebo heuristických metód.

Jazykovedný ústav sprístupňuje rozhranie, ktoré demonštruje lematizáciu, morfologickú anotáciu a dezambiguáciu slovenských textov na adrese https://morphodita.juls.savba.sk/ (aktuálna verzia k 6. 4. 2021).

Analýza textu je založená na morfologickej databáze, ktorá obsahuje 111-tisíc lem; 3,6 milióna záznamov; 1,3 milióna jedinečných slovných tvarov. Samotná

[1] Výnimkou je v našom prípade napr. lematizácia aglutinovaných tvarov zámen, napríklad *doňho* má priradenú lemu *do_on*, a lematizácia interpunkčných znakov.

[2] Toto neplatí napríklad pre polysyntetické jazyky, kde nie je ani takéto „klasické základné“ počítačové spracovanie doteraz uspokojivo vyriešené.

[3] Ak sa priraďuje tento význam k podstatným menám ako substantivizovaná vlastnosť v skupine adjektiválií (Sokolová, 2007). Tu používame „klasické“ slovnodruhové označenie, miešajúce morfologickú a syntaktickú rolu. Za zmienku stojí, že de facto štandardný morfologický tagset používaný vo významných slovenských korpusoch elegantne vyriešil túto problematiku zavedením *paradigmy* ako dodatku k *slovnému druhu*, konkrétne v tomto prípade ide o substantíva s adjektívnou paradigmou, čo je plne reflektované aj v popisovanom rozhraní. Samozrejme, už menej elegantne pôsobí „presnosť“ automatickej dezambgiuácie takýchto často prakticky neodlíšiteľných kategórií.

analýza a dezambiguácia používa softvér MorphoDita[4] vrátane jednoduchého štatisticko-heuristického guessera[5], používaného na odhad možnej lemy a gramatických kategórií slova, ktoré sa nenachádza v morfologickej databáze (takéto slová sú v analyzovanom výstupe označené špeciálnym reťazcom «*Toto slovo nepoznám*»).

V rozhraní je možné zadať krátky text (v rozsahu niekoľkých odsekov), ktorý bude automaticky lematizovaný a morfologicky označkovaný (v tomto procese je text aj tokenizovaný[6] a segmentovaný na vety). Okrem zadania vlastného textu je možné zadať aj voľbu *Ukážka*, ktorá náhodne vyberie niekoľko viet slovenského textu[7] na účel ukážky, bez potreby zadávania vlastného textu.

Vo výsledku sú jednotlivé vety zobrazené vo forme riadkov, kde pod každým slovom je jeho lema, a gramatické (morfologické) kategórie, resp. ich hodnoty sa zobrazia, keď používateľ presunie kurzor myši ponad dané slovo. Anotácia gramatických kategórií je založená na morfologických značkách, ktoré sa používajú v hlavných slovenských korpusoch. Tu je navyše zobrazená vo forme krátkej vety, ktorá približuje gramatické kategórie daného slova nevtieravým a prirodzeným spôsobom.

Rozhranie je dostupné v niekoľkých jazykoch[8], slovenčina v dvoch variantoch – «*odborná slovenčina*» a «*laická slovenčina*». Tieto verzie sa líšia použitou terminológiou. *Odborná slovenčina* používa väčšinou pôvodom latinské lingvistické termíny (*substantívum, verbum, adjektívum* atď.), *laická slovenčina* pracuje so slovenskými termínmi (*podstatné meno, sloveso, prídavné meno*…) a niektoré informácie sú pri nej opísané zjednodušeným spôsobom.

Na ilustráciu funkcie uvedeného rozhrania uvádzame ukážky rozhrania v odbornej slovenčine, laickej slovenčine a esperante (obrázky 1, 2, 3). Vo vrchnej časti je zobrazený analyzovaný text, kde je každá (v tomto prípade jedna) veta v samostatnom riadku. Pod jednotlivými slovami sa zobrazujú ich lemy. V ďalšom riadku sa zobrazuje informácia o gramatických kategóriách slova, na ktorom sa práve nachádza kurzor myši (v tomto prípade „*nedarit*“). Ďalej nasleduje textové pole určené na zadávanie textu na analýzu a ovládacie prvky – spustenie analýzy (*Analyzuj*), analýza niekoľkých náhodne vybraných viet (*Ukážka*) a výber jazyka rozhrania.

[4] Dostupný na: http://ufal.mff.cuni.cz/morphodita.

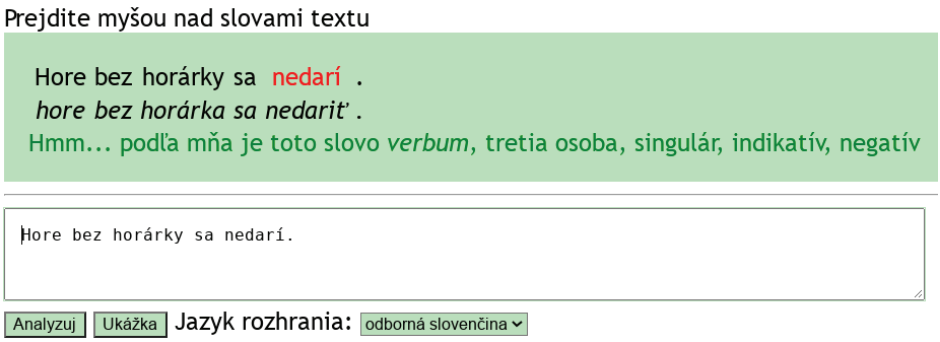
[5] Algoritmus, ktorý sa snaží uhádnuť lemu, slovný druh a hodnoty gramatických kategórií slov, neobsiahnutých v morfologickom slovníku, v našom prípade na základe štatistických vlastností sufixu a prefixu daného slova a jeho kontextu. Nebudeme sa snažiť vymýšľať slovenský termín, ale vypomôžeme si prevzatým slovom *guesser*.

[6] Tokenizácia je rozdelenie textu na základné (korpusové) jednotky – tokeny – zodpovedajúce slovám, znakom interpunkcie, číslicam a podobne.

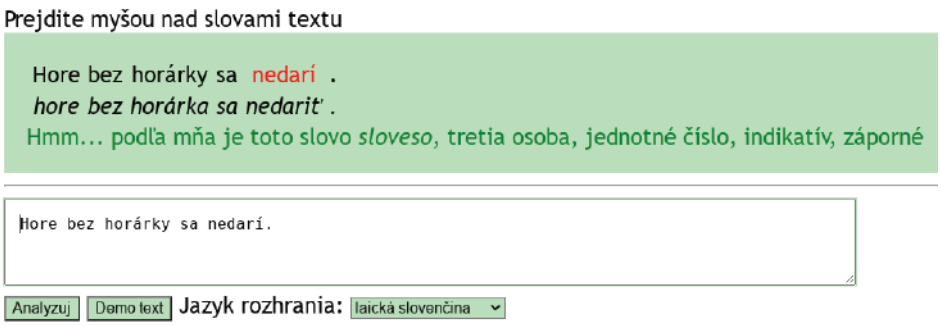
[7] Ide o výber viet z niekoľkých slovenských odborných a literárnych diel.

[8] V slovenčine, angličtine, esperante, ruštine, litovčine, francúzštine a nemčine[9].

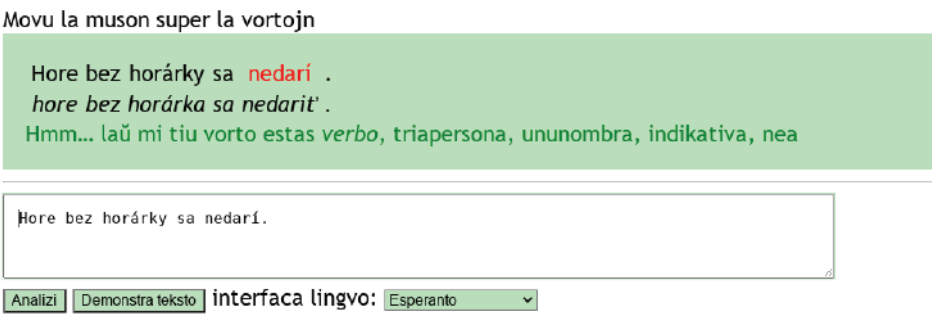
[9] Ide samozrejme iba o jazyky rozhrania. Lematizácia a morfologická analýza prebieha iba na textoch v slovenskom jazyku.



Obr. 1. Ukážka rozhrania, jazyk «*odborná slovenčina*».



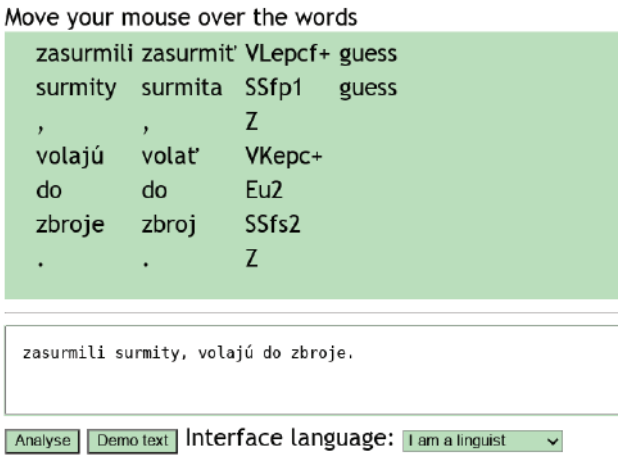
Obr. 2. Ukážka rozhrania, jazyk «*laická slovenčina*».



Obr. 3. Ukážka rozhrania, jazyk «*esperanto*».

Okrem toho rozhranie poskytuje špecifickú voľbu vyhľadávania (v položke *Jazyk rozhrania*) prostredníctvom možnosti nazvanej «*I am a linguist*». Ak používateľ

zvolí túto možnosť, jeho vyhľadávanie sa prepne do režimu, v ktorom morfologické značky nebudú prekladané do bežného jazyka a zobrazenie analyzovaných viet bude mať formu vertikálneho súboru, aký je známy napríklad z korpusového spracovania jazyka (pozri obrázok 4). Tento režim bol primárne vytvorený na pedagogické účely, ale svoje uplatnenie nachádza medzi poučenými používateľmi, ktorí poznajú štruktúru použitých morfologických značiek. Štatisticko-heuristické hádanie lemy, slovného druhu a hodnôt gramatických kategórií (morfosyntaktickej značky) je zaznačené prítomnosťou špeciálneho reťazca znakov «*guess*» v štvrtom stĺpci výstupu, na obr. 4 je demonštrované pri dvoch neznámych slovách *zasurmili* a *surmity*, kde guesser správne určil hodnoty gramatických kategórií a lemy, iba v prípade slovesného vidu sa mierne pomýlil (nedokonavý namiesto dokonavého).



Obr. 4. Ukážka rozhrania, voľba «*I am a linguist*».

Vo všeobecnosti poskytované nástroje zameriavajúce sa na počítačové spracovanie prirodzeného jazyka nachádzajú uplatnenie (často neviditeľné) aj medzi bežnými používateľmi jazyka. Okrem praktického aplikovaného použitia a využitia pri vedeckom výskume majú tieto nástroje potenciál slúžiť ako efektívne prostriedky pri výučbe slovenčiny ako cudzieho jazyka. Zmyslom predstaveného nástroja je priblížiť proces lematizácie a dezambiguácie, predovšetkým ich výsledok, ktorý je zrozumiteľný na laickej úrovni, ale poskytuje aj výsledky na expertnej úrovni pre hlbšie morfologické analýzy jazyka (jazykov). Lematizácia, morfologická anotácia a dezambiguácia tvoria základ ďalšieho počítačového spracovania jazyka. V oblasti jazykových technológií predstavujú veľmi dôležitú súčasť činnosti Jazykovedného ústavu Ľ. Štúra SAV, ktorú prostredníctvom rozhrania k slovenskému modelu pre značkovací systém MorphoDiTa (Straková et al., 2014) približujeme nielen odbor-

nej, ale aj širokej verejnosti, ktorá sa zaujíma o problematiku počítačového spracovania slovenského jazyka.

Literatúra

BENKO, Vladimír: Aranea: Yet Another Family of (Comparable) Web Corpora. In: Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Eds. P. Sojka – A. Horák – I. Kopeček et al. Springer International Publishing Switzerland, 2014. s. 257 – 264.

GARABÍK, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference, Bratislava 2006, s. 2 – 5.

SOKOLOVÁ, Miloslava: Nový deklinačný systém slovenských substantív. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2007. 345 s.

STRAKOVÁ, Jana – STRAKA, Milan – HAJIČ, Jan: Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland. 2014. s. 13 – 18.

Slovenský národný korpus – prim-*.0-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. Dostupný na: https://korpus.juls.savba.sk.