SLOVENČINA A ČEŠTINA V POČÍTAČOVOM SPRACOVANÍ

Zborník referátov zo seminára, Bratislava 26. – 27. októbra 2001,

Veda, Vydavateľstvo Slovenskej akadémie vied, Bratislava 2001, 194 s.

Článok bol publikovaný v Jazykovednom časopise (2003, ročník 54, č. 1-2, s. 91 – 95).

Zdá sa, že problematika počítačovej a korpusovej lingvistiky sa posúva čoraz viac do stredobodu pozornosti odborných lingvistických a informatických sfér aj v našich zemepisných šírkach. Prejavom toho bola aj konferencia o počítačovom spracovaní slovenčiny a češtiny, ktorá sa uskutočnila v Bratislave polovicou jesene roku 2001. Zámerom organizátorov bolo zvýšenie vzájomnej informovanosti medzi slovenskými odbornými kruhmi zaoberajúcimi sa počítačovou lingvistikou, ale aj prezentácia výsledkov práce v počítačovom spracovaní češtiny prostredníctvom vedúcich českých odborníkov z tejto oblasti. O interdisciplinarite počítačového spracovania prirodzeného jazyka svedčí aj široký záber a polia pôsobnosti v ktorých pracujú účastníci tejto konferencie. Príspevky z nej vyšli v zborníku, ktorý je predmetom prítomnej recenzie.

Zborník otvára štúdia Jana Hajiča (Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta UK, Praha) *Statistické modelování a automatická analýza přirozeného jazyka* (s. 11 – 33). Jej názov do istej miery odzrkadľuje aj prevládajúci trend v súčasnej počítačovej lingvistike. Štatistické a pravdepodobnostné modelovanie má v súčasnosti mnoho praktických aplikácií, z ktorých známejšie sú napr. OCR (Optical Character Recognition), čiže optické rozpoznávanie naskenovaného textu, alebo analýza a syntéza akustického rečového signálu. Hajič sa v príspevku zameriava na automatickú morfologickú analýzu a dezambiguáciu, jej prepojenie na (polo-) automatickú syntaktickú analýzu a automatický preklad medzi geneticky príbuznými jazykmi. V procese automatickej morfologickej analýzy je východiskom morfologický slovník, ktorý obsahuje informácie o prípustných lexikálnych a relačných morfémach v danom jazyku, ďalej súbor značiek (angl. tagset) , ktoré reprezentujú morfologické kategórie a matematický model, podľa ktorého analýza prebieha. Nevyhnutným predpokladom je aj tzv. trénovací korpus, čiže súbor textov, ktorý bol predtým ručne morfologicky analyzovaný podľa vopred zvoleného tagsetu. Na tomto korpuse sa program, ktorý je implementáciou daného štatistického modelu, "naučí" pravdepodobnostné vzťahy medzi slovnými tvarmi a ich možnými morfologickými interpretáciami. Keďže je výsledok takejto analýzy v dôsledku tvarovej homonymie pomerne nejednoznačný (napr. politika N. sg. ž. r. aj G. sg. m. r.), aplikuje sa proces dezambiguácie, ktorý takisto pracuje na

štatistickom základe. Úspešnosť takejto analýzy sa pre češtinu a iné slovanské jazyky pohybuje okolo 94 %. Automatickú morfologickú analýzu Hajič využíva aj v strojovom preklade vo fáze analýzy a syntézy slovných tvarov východiskového a cieľového jazyka. Výsledky strojového prekladu medzi geneticky (a štruktúrne) príbuznými slovanskými jazykmi sú pomerne dobré pre slovenčinu a češtinu (okolo 90%), horšie pre češtinu a poľštinu (asi 75%).

V príspevku *Identifikácia paradigmatických a syntagmatických vzťahov v texte* (s. 34 – 48) sa Karol Furdík (Juvier, s. r. o, Košice) venuje popisu teoretického modelu, ktorý by bol schopný identifikovať a reprezentovať paradigmatické a syntagmatické vzťahy v texte. Tento model je založený na algoritme induktívneho inkrementálneho učenia. Takýto algoritmus by pozostával z viacerých modulov, ktoré by vedeli vstupný text segmentovať, identifikovať a explicitne reprezentovať jazykové jednotky pomocou vhodne zvoleného formálneho jazyka.

Článok Eduarda Kostolanského a Jany Hašanovej (Laboratórium počítačovej lingvistiky, Pedagogická fakulta UK, Bratislava) *Spracovanie morfologickej roviny slovenčiny počítačom* (s. 49 – 87) predstavuje opis modelu morfologickej analýzy a generovania slovných tvarov v slovenčine. Hlavnou myšlienkou je možnosť generovania slovných tvarov slovenčiny pomocou virtuálneho slovníka, ktorý nesie komplexné údaje o vzoroch ohýbania, tvarotvorných základoch a ich zmenách, súbore ohýbacích prípon, pravidlách formovania jednotlivých slovných tvarov a údaje o konkrétnych ohýbaných slovách. Všetky tieto údaje sa získali z textového materiálu Krátkeho slovníka slovenského jazyka a boli uložené vo forme referenčnej morfologickej databázy.

Výskumný tím Milan Rusko – Sachia Daržágín – Marián Trnka z Ústavu informatiky SAV prezentuje výsledky svojho výskumu v článku *Databázy pre rozpoznávanie a analýzu reči v slovenčine* (s. 88 – 97). Takéto databázy majú v komerčnej sfére aplikácie v hlasových telefonických službách, ktoré poznáme z každodenného života (napr. informácie o odchode dopravných spojení, teleshopping, telebanking, voice mail atď.). Autori sa zamýšľajú nad súčasnými teoretickými prístupmi v rozpoznávaní reči (skryté markovovské modely a neurónové siete) a predstavujú priebeh a výsledok tvorby prvej slovenskej telefonickej rečovej databázy SpeechDat-e.

Jozef Ivanecký (IBM Voice Systems, European Speech Research; Katedra kybernetiky a umelej inteligencie , Fakulta elektrotechniky a informatiky, Technická univerzita, Košice) sa v dvoch príspevkoch venuje automatickej fonetickej transkripcii v počítačovom rozpoznávaní reči. V prvom, nazvanom *SAMPA v slovenčine a jej význam z pohľadu viacjazyčných systémov na rozpoznávanie reči* (s. 98 – 116), autor prezentuje vlastné skúsenosti pri vývine

fonetickej abecedy SAMPA (Speech Assesment Methods – Phonetic Alphabet) pre slovenčinu. Vo viacjazyčných systémoch na rozpoznávanie reči je hlavnou požiadavkou vzájomná kompatibilita jednotlivých transkripčných systémov. Podľa Ivaneckého sú výhodami kódovania SAMPA práve vhodnosť a rozšírenosť použitia vo viacjazyčných systémoch. V druhom príspevku s názvom *Automatická transkripcia slovenčiny v počítačovom rozpoznávaní reči* (s.117 – 120) sa Ivanecký venuje problematike automatickej fonetickej transkripcie v teoretickej rovine a predstavuje možný spôsob riešenia tohto problému v slovenčine. Pre slovenčinu volí autor transkripčné pravidlá vychádzajúce z prác A. Kráľa, výstupnou abecedou je autorom navrhovaná SAMPA. Ako problematický jav sa Ivaneckému ukázal rozdiel v spisovnej a hovorovej výslovnosti. Keďže systém pracuje primárne s pravidlami pre spisovnú výslovnosť, bolo nutné doňho implementovať aj pravidlá pre nespisovnú výslovnosť, pretože ich zanedbanie by mohlo viesť k chybným výstupom v transkripcii.

Ukážkou, že peniaze európskych daňových poplatníkov sa prostredníctvom európskych inštitúcii môžu niekedy prerozdeľovať aj zmysluplne, je projekt COP-58 ONOMASTICA, ktorý vznikol v rámci programu COPERNICUS. Jeho výsledky prezentuje slovenský účastník projektu Peter Ďurčo (Akadémia policajného zboru, Bratislava) v príspevku *Viacjazyčný výslovnostný slovník vlastných mien a názvov lokalít.* V rámci tohto projektu boli vytvorené jazykové zdroje obsahujúce okolo 1,8 milióna vlastných mien v siedmich jazykoch strednej a východnej Európy (čeština, slovenčina, poľština, estónčina, litovčina a ukrajinčina). Autor sa v príspevku venuje technickým aspektom (kódovanie textov, používateľské rozhranie) využitia týchto textových databáz a poukazuje (ak nie rovno nabáda) na možnosť použitia týchto hodnotných dát na ďalší výskum v počítačovej a korpusovej lingvistike.

Dva príspevky pražských korpusových lingvistov z Ústavu Českého národného korpusu sú prezentáciou aktuálneho stavu zatiaľ najrozsiahlejšieho a úrovňou spracovania najkvalitnejšieho korpusu textov v slovanských jazykoch. V teoreticko-metodologickej štúdii *Český národní korpus: Stav v roce 2001* (s. 121 – 135) sa vedúci českého korpusového projektu, František Čermák, vracia k základným pojmom z oblasti korpusovej lingvistiky (korpus, reprezentatívnosť, štandardizácia, značkovanie, konkordancie), predstavuje fázy budovania a súčasnú štruktúru Českého národného korpusu a naznačuje aspekty práce a využitia korpusových dát v lingvistickej ako aj nelingvistickej sfére. Michal Křen sa zase v príspevku *Český národní korpus – počítačová demonstrace* (s. 136 – 141) zameriava na popis technických aspektov budovania a využitia ČNK (získavanie textov, skenovanie, prevod textov do jednotného formátu, externá anotácia, používanie korpusového manažéra) a

problémom s tým spojeným.

Vlasta Křečková (Fakulta humanitných vied UMB, Banská Bystrica) v príspevku *Informácie v terminologických databázach a ich využívanie pri preklade odborných textov* (s. 142 – 149) predstavuje aspekty tvorby terminologických databáz z rozličných oblastí (sociálna ochrana, vodohospodárstvo, ochrana životného prostredia, knižničná a informačná veda) v slovenskom aj európskom kontexte. Křečková popisuje štruktúru položiek terminologickej databázy, ktoré môžu niesť lingvistické aj encyklopedické informácie. Zvyčajne terminologický záznam v lingvistickej rovine obsahuje okrem termínu v základnej gramatickej forme aj jeho synonymá, varianty, symboly, skrátené formy, inojazyčné ekvivalenty, kolokácie, grafické reprezentácie, hyponymá a hyperonymá. Ďalej sú tu technické informácie o identifikátore termínu, symbole jazyka, dátume vytvorenia záznamu, identifikátori tvorcu záznamu, kóde inštitúcie, informáciach o zdrojoch a pod. Autorka propaguje myšlienku využitia (a pravdaže aj budovania) takto informačne bohatej databázy v prekladateľskej činnosti a podáva stručnú sondu o používateľských návykoch prekladateľov s terminologickými databázami (najčastejšie vyhľadávané položky a pod.).

Ojedinelým a pozoruhodným je projekt slovníka štúrovskej slovenčiny a jeho počítačovej podpory, ktorý vo svojom rovnomennom článku (s. 150 – 154) predstavuje Ľubomír Kralčák (Katedra slovenského jazyka Filozofickej fakulty UKF Nitra). Ústredným problémom v tomto projekte je podľa autora otázka periodizácie textov štúrovského obdobia, ktorú nazýva vertikálnou rovinou vymedzenia inventára textov. V záujme zachytenia vzácnych jazykových javov sa prihovára za jej maximalizáciu. Hornú a dolnú časovú hranicu pre výber štúrovských textov zatiaľ nekonkretizuje, iba uvádza rozdielne chápanie štúrovského obdobia u viacerých jazykovedcov (Pauliny, Furdík). Ďalšími problematickými okruhmi sú otázky vhodnosti zaradenia textov podľa kritéria publikovanosti/rukopisnosti textového materiálu a lexikálne okazionalizmy, kalky a výpožičky štúrovských autorov. Z informatického hľadiska podpory projektu Kralčák uvádza naliehavosť vytvorenia efektívneho vyhľadávacieho softwaru a cieľ vybudovať textový korpus štúrovčiny v rozsahu 15 miliónov slovných tvarov.

Nevyhnutnosť existencie nového reprezentatívneho slovníka súčasnej češtiny zdôrazňuje Karel Pala (Katedra informačních technologií, Fakulta informatiky, MU Brno). Jazykovou materiálovou základňou pre jeho vznik by mala byť Česká lexikálna databáza (CLD), ktorej koncepciu Pala predstavuje v príspevku *Návrh české lexikální databáze* (s. 155 – 167). Databáza by mala obsahovať približne 60 000 heslových slov a 105 000 významov. Z tohto počtu až 20 000 jednotiek by tvorili slovesá, ktoré Pala považuje za základné relačné prvky v jazyku, viažuce na seba iné slovné druhy (hlavne substantíva). Základnými jednotkami

v databáze by mali byť lémy (slovo v základnom, "slovníkovom" gramatickom tvare), resp. štandardné kolokácie (napr. *telefónny zoznam, vysoká škola* a pod.). Lingvistické informácie , ktoré sa v databáze podľa návrhu majú ocitnúť sú: fonologická (fonetická) informácia, morfologická informácia, významy, sémantické rysy, hypero/hyponymické vzťahy, sémantické triedy, synonymické rady, syntaktické informácie, lokálne kontexty, typické použitie, kolokácie, pragmatické informácie, pôvod, logický typ a encyklopedická informácie. Tieto položky sa v databáze reprezentujú ako polia, zapísané vo formáte XML. Na konci príspevku sa uvádzajú aj konkrétne príklady možných slovných záznamov v databáze.

Na Palov príspevok obsahovo nadväzuje Pavel Smrž (Fakulta informatiky, MU Brno) v článku *Slovníková data ve formátu XML* (s. 168 – 180). Zdôrazňuje výhody použitia značkovacieho jazyka XML (eXtensible Markup Language) na reprezentáciu zložito štruktúrovaných dát, akými sú práve slovníkové dáta. Medzi tie by mali patriť schopnosť validácie formy a obsahu dokumentu pomocou definičného jazyka XML Schema, možnosť transformácie dokumentu a jeho elementov do ľubovoľného formátu pomocou formátovacieho jazyka XSLT (eXtensible Stylesheet Language), výkonný mechanizmus zadávania vyhľadávacích otázok XQuery, štandardy pre definíciu odkazov medzi XML dokumentmi mechanizmom Xlink atď. Okrem zdôrazňovania výhod XML uvádza Smrž ako najznámeší projekt slovníkového využitia XML anglický výkladový slovník Oxford English Dictionary Online a popisuje aj priebeh prevodu Slovníka spisovného slovníka jazyka českého do XML formátu. Získané skúsenosti formuluje ako odporúčania pre obdobné projekty.

Zborník uzatvára retrospektívny pohľad Vladimíra Benka (Laboratórium počítačovej lingvistiky, Pedagogická fakulta UK, Bratislava) na počítačovú podporu lexikografických projektov na Slovensku (s. 181 – 194). Autor spomína na začiatky svojej práce v oblasti elektronizácie slovníkových dát, ktorá sa ukázala byť vo svojom čase (1987) nevyhnutnou (a poslednou) možnosťou záchrany 1. vydania Krátkeho slovníka slovenského jazyka. V priebehu jej uskutočňovania Benko navrhol značkovací jazyk MOM (My Own Markup), ktorým sa zaznamenávajú informácie o štruktúre slovníka z typografického hľadiska (kurzíva, (polo)tučné písmo, rôzne veľkosti písma a pod.). Táto forma slovníkovej reprezentácie sa ukázala byť natoľko efektívna z hľadiska práce lexikografov i kvality slovníka, že sa uplatnila aj v 3. a 4. zväzku Historického slovníka slovenského jazyka (1992) a (1993), Synonymickom slovníku slovenčiny (1995) a 3. vydaní Krátkeho slovníka slovenského jazyka (1997).

Alexander Horák