Hovorený korpus slovenčiny

Mária Šimková – Radovan Garabík – Agáta Karčová – Katarína Gajdošová

1. Slovenský národný korpus

Projekt Slovenský hovorený korpus (***Hovor***) je súčasťou Slovenského národného korpusu (http://korpus.juls.savba.sk; ďalej SNK). SNK sa začal systematicky budovať v Jazykovednom ústave Ľ. Štúra SAV v Bratislave v r. 2002 a v prvej etape (do r. 2006) sa jeho budovanie sústreďovalo na písané texty súčasného slovenského jazyka od r. 1955 do r. 2005, čo je v korešpondencii s obdobím spracúvaným v novom výkladovom slovníku slovenského jazyka, ktorý sa koncipuje predovšetkým na báze korpusového materiálu. Začiatkom r. 2007 bola ako výsledok prvej fázy sprístupnená na Internete najnovšia, šiesta verzia hlavného, základného korpusu ***prim-3.0*** v rozsahu 350 miliónov textových jednotiek a druhá verzia ručne morfologicky anotovaného korpusu v rozsahu takmer 512 tisíc textových jednotiek. Každý text v korpuse je podložený súhlasom autora alebo majiteľa autorských či distribučných práv na jeho spracovanie a zaradenie do celku korpusu podľa licenčnej zmluvy a má podrobnú bibliografickú a štýlovo-žánrovú anotáciu. Celý korpus je automaticky lematizovaný a automaticky morfologicky označkovaný po natrénovaní značkovacieho softvéru na ručne morfologicky anotovaných textoch. Vybrané texty sa ručne anotujú aj syntakticky. Okrem hlavného, jednojazyčného korpusu sa pracovalo aj na tvorbe paralelných korpusov, z ktorých sú sprístupnené zatiaľ tri (Parallel Corpus of Computer Terms, Francúzsko-slovenský paralelný korpus, Rusko-slovenský paralelný korpus), pripravujú sa ďalšie (slovensko-český, slovensko-chorvátsky, slovensko-nemecký a slovensko-anglický paralelný korpus).

V druhej etape by sa mal Slovenský národný korpus rozrásť na 600 miliónov textových jednotiek v základnom korpuse písaných textov od r. 1955 do súčasnosti, t. j. do r. 2011, dokedy bola druhá časť projektu schválená. Osobitnou súčasťou celého projektu Slovenského národného korpusu je v druhej etape tvorba Slovenskej terminologickej databázy (https://data.juls.savba.sk/std/) a ***budovanie korpusu hovorenej slovenčiny***, ktorej výskum je aktuálne jednou z dôležitých úloh viacerých slovakistických pracovísk na Slovensku (Prešov, Banská Bystrica, Nitra, Bratislava).

2. Hovorená slovenčina – vymedzenie a doterajšie výskumy

Stratifikácia slovenčiny a skúmanie jej nepísanej (a nenárečovej) formy má svoje vlastné špecifiká a tradíciu. Podľa J. Horeckého (1979, 1984) sa v súčasnej slovenčine vymedzuje popri spisovnej forme a teritoriálnych a sociálnych formách aj štandardná a subštandardná forma s celoslovenskou platnosťou, J. Bosák (1990, 1993) používa termín hovorová komunikačná varieta slovenčiny. Výskum hovorenej komunikácie na východnom Slovensku (Slančová, Sokolová 1995) priviedol autorky k záveru, že

> okrem celoslovenskej podoby týchto foriem slovenčiny, štandardnej formy slovenčiny (bežne hovorenej slovenčiny) možno predpokladať i existenciu jej regionálnych variantov ..., utvárajúcich sa a fungujúcich na západnom, strednom a východnom Slovensku, a o lokálnych podobách týchto variantov, fungujúcich hlavne v komunikácii miest (133).

Tieto závery korešpondujú s výsledkami materiálového výskumu slovenčiny v Banskej Bystrici, kde V. Patráš (1990) registroval: ústnu podobu spisovnej slovenčiny, formujúcu sa hovorovú slovenčinu, dialekty a sociolekty, češtinu a zložky iných jazykových sústav a expanzívne prejavy idiolektov. Na konferencii *Mesto a jeho jazyk* (Bratislava 1998) sa konštatovalo, že podobné výskumy sú na Slovensku veľmi ojedinelé a mal by sa rozvinúť celoplošný materiálový výskum hovorenej slovenčiny v podobnom rozsahu, ako bol zorganizovaný začiatkom 60. rokov 20. st. (*Hovorená podoba spisovnej slovenčiny*, 6.-9. októbra 1965). Žiaľ, výsledky tohto výskumu ostali vo veľkej miere nevyužité, a tak existujúca ortoepická norma slovenčiny (Kráľ 2005) nevychádza zo systematického materiálového výskumu reálnej komunikačnej situácie, ktorá sa posledných 20-30 rokov dynamicky mení. Nové poznatky môžu priniesť výskumy aktuálne realizované pracovníkmi Prešovskej univerzity v Prešove (detská reč; žurnalistické reportáže), Univerzity Mateja Bela v Banskej Bystrici (Dynamika spoločenských zmien a stratifikácia slovenčiny), Univerzity Konštantína Filozofa v Nitre (Hovorená podoba spisovnej slovenčiny v masmédiách – so zameraním na spontánne prehovory), ako aj začínajúci projekt Slovenského hovoreného korpusu v Jazykovednom ústave Ľ. Štúra SAV v Bratislave. Ako reálne sa ukazuje využitie výsledkov a skúseností doterajších výskumov v oblasti analýzy a syntézy reči (najmä Ústav informatiky SAV v Bratislave).

3. Základné východiská tvorby Slovenského hovoreného korpusu

3.1. Cieľ korpusu

Slovenský hovorený korpus sa koncipuje ako všeobecná elektronická databáza hovorených komunikátov s čo najširším využitím:

1) Materiál na opis hovorenej podoby slovenčiny (okrem ortoepických aj lexikálne a gramatické charakteristiky)

2) Báza bežnej i špecifickej (tabuizované slová) hovorenej lexiky pre potreby lexikografického spracovania

3) Materiál na rečové analýzy

4) Elektronicky spracovateľný a použiteľný materiál

5) Sprístupnenosť na vedecko-výskumné využitie na Internete

3.2. Štruktúra korpusu

Predpokladá sa:

1. Veľkosť 2 milióny textových jednotiek
2. Rovnomerné zastúpenie respondentov podľa

a) demografických ukazovateľov

pohlavie (žena – muž)

vek (20 – 30, 31 – 40, 41 – 50, 51 – 60, 61 – 70, 71 – 80)

vzdelanie (nižšie, stredoškolské, vysokoškolské)

b) geografických ukazovateľov

miesto narodenia

miesto najdlhšieho pobytu

miesto súčasného pobytu/pôsobenia

a zohľadňovanie

c) sociologických a pragmatických faktorov

ne/poznanie, ne/rovnocennosť partnerov, ne/formálnosť komunikácie

d) lingvistických faktorov

monológ – dialóg – trialóg – ...

riadený – spontánny rozhovor

3. Špecifické zásady nahrávania, resp. možnosti získavania nahrávok, a prepisu, ako aj úsilie o dodržanie právnych predpisov a etických noriem povedú zrejme k existencii viacerých podkorpusov:

a) autorizované – neautorizované

s plnou – s čiastočnou anotáciou relevantných údajov o respondentoch

b) prístupné verejne – prístupné interne,

c) prepísané ortograficky s cielenou transkripciou – malý podkorpus s podrobnou transkripciou.

3.3. Zásady nahrávania, technického spracovania a uchovávania nahrávok

Nahrávanie pracovníkmi Slovenského národného korpusu sa bude realizovať pomocou špeciálneho digitálneho záznamníka, z iných zdrojov sa budú získavať (resp. už existujú) nahrávky na bežných nosičoch (pásové a kazetové magnetofóny). Záznamy spontánnych prehovorov z médií budú okrem zvuku obsahovať aj obraz, ktorý sa perspektívne môže začleniť do systému korpusu ako ďalšia úroveň záznamu.

Vzhľadom na prístupné ceny záznamových médií v súčasnosti je realizovateľné uchovávať digitálne zvukové záznamy so zachovaním všetkých informácií, z čoho vyplýva potreba použitia bezstratových formátov záznamu. Pre potreby SHK sme vybrali FLAC (http://flac.sf.net/) lossless audio codec, ktorý dosahuje kompresné pomery pre bežné zvukové súbory na úrovni približne 50 %. Časť súborov však bude v rôznych iných formátoch (stratových) tak, ako sa nám ich podarí získať.

Na prezentáciu hovorenej časti korpusu sa ako najvhodnejší javí SPEEX codec (http://www.speex.org/), prístupné úseky sú downmixované do jednokanálového zvuku so vzorkovacou frekvenciou 32 kHz a kódované v kvalite 6 (priemerný bitrate 23 kb/s). Ako alternatívu k menej rozšírenému SPEEX codecu poskytneme aj Ogg/Vorbis formát. Do Vorbis codecu sme kódovali jednokanálový zvuk v kvalite -1, čo predstavuje priemerný bitrate asi 40 kb/s, s použitím experimentálnych aoTuV patchov, optimalizovaných pre nízke bitrates. Okrem toho sa plánuje aj Java applet prehrávajúci SPEEX codec pre bežných (po)užívateľov. Najrozšírenejší codec MP3 bol vylúčený z právnych a finančných dôvodov. V prípade záujmu o počúvanie zvukového záznamu nami vytvoreného korpusu bude potrebné, aby si každý jednotlivý používateľ nainštaloval niektorý z prístupných codecov alebo použil javový prehrávač.

Slovenský hovorený korpus bude okrem grafického prepisu využiteľný aj prostredníctvom zvukových záznamov zlinkovaných navzájom s prepísaným textom.

3.4. Zásady prepisu

Existujú viaceré možnosti prepisu hovorených komunikátov do grafickej podoby, pričom každá z možností nesie so sebou pozitíva, ale aj negatíva, ako ukazujú doterajšie skúsenosti z práce s hovorenými korpusmi (napr. Čermák 2006, Esvan 2006, Pořízka 2004, 2005).

Fonetický prepis je najpodrobnejší prepis s najnižším stupňom abstrakcie, zohľadňuje varianty a zvukové odtienky každej hlásky. Fonetická transkripcia zachytáva hovorenú podobu jazyka najpresnejšie a najdôkladnejšie, ale vzhľadom na plánovaný rozsah korpusu hovorených komunikátov (cca 2 milióny tokenov) je pre časovú a odbornú náročnosť nepoužiteľný. Takisto by takýto spôsob prepisu neprispieval k prehľadnosti textu a prihliadajúc na dôležité, aj keď nie jediné účely využitia korpusu hovorených komunikátov – na gramaticko-lexikálno-štylistické a sociolingvistické výskumy, by bol nepraktický a nepoužiteľný na sprístupnenie pomocou doteraz vyvinutých nástrojov na prácu s písanými korpusmi. V ďalších fázach projektu sa však nevylučuje možnosť utvorenia podkorpusu s podrobným fonetickým prepisom.

Vzhľadom na naše technické možnosti máme v zásade na výber tri spôsoby transkripcie nahratých zvukových záznamov do grafickej podoby:

1) podobne ako v českých hovorených korpusoch (Pražský mluvený korpus, Brněnský mluvený korpus a ORAL 2006) prepisovať štandardným pravopisom (ktorý je kombináciou fonetického a etymologického zápisu), ale ak sa výslovnosť odlíši od štandardu, prepísať slovo tak, ako bolo reálne vyslovené (napr. *ďiplomat*); problémom zostáva, do akej miery by sa zachytávala znelostná asimilácia;

2) možno prepisovať štandardným pravopisom a v prípade, že sa reálna výslovnosť líši od štandardnej výslovnosti, zapísať okrem pravopisnej aj výslovnostnú verziu (napr. do zátvoriek alebo iným dohodnutým znakom na vyčlenenie tohto zápisu):

napr. «ukáž takú vetu kde#gde sa to ukáže»
alebo
«ukáž#ukáš takú vetu kde#gde sa to ukáže»

3) možno prepisovať štandardným pravopisom a všetko, čo je vo výslovnosti odlišné od písanej podoby, zapisovať osobitným zápisom vyčleneným zátvorkami alebo iným dohodnutým znakom:

napr. «ukáž#ukaš takú vetu kde#gďe sa to ukáže»
(pozn. ukaš s krátkym a sa zapíše v prípade, že *a* bolo vyslovené krátko).

Nevýhodou posledných dvoch spôsobov zápisu je zdĺhavosť a rozsah, výhodou aplikácia automatických nástrojov na analýzu.

Je teda dôležité nájsť rozumný kompromis medzi fonologickým a tradičným zápisom podľa štandardnej pravopisnej normy tak, aby bol text prepisu prehľadný a zrozumiteľný, ale aby zároveň zachytával čo najvernejšie špecifiká konkrétnych prehovorov.

Prehovorom rozumieme súvislý neprerušovaný prejav jedného hovoriaceho respondenta (účastníka dialógu). V koncepcii SHK sa smeruje k zapisovaniu celého prejavu všetkých respondentov vrátane otázok pýtajúceho sa (riadiaceho člena rozhovoru) aj vzhľadom na to, že otázky budú rôzne, resp. vstupy nahrávajúceho budú pôsobiť smerom k čo najväčšej spontánnosti respondenta. Začiatok a koniec každého prehovoru oddelíme špeciálnym znakom. Čiastkové zachytenie melódie vety a páuz sa bude realizovať použitím tradičných interpunkčných znamienok na konci viet (otáznik, bodka, výkričník) a čiarok podľa pravopisnej normy s doplnením zachytenia reálne sa vyskytujúcich páuz rôznej dĺžky a druhu (pauzy fyziologické, logické, pauzy na neočakávaných miestach, tzv. dramatické a iné) bez ohľadu na štruktúru a prehľadné členenie výpovede. V prípade kombinácie zápisov páuz sa zrejme pristúpi ku kompromisu ponechaním otáznika ako znaku signalizujúceho koniec vety a zároveň antikadenciu, príp. semikadenciu, ostatné pauzy sa budú zapisovať prehľadným znakom – spojovníkom – jedným v prípade krátkej pauzy, dvoma pri stredne dlhej pauze a troma pri veľmi dlhej pauze.

Prípadný prepis ďalších suprasegmentálnych javov, ako sú melódia vety, emfáza, prípadne aj slovný a vetný prízvuk a ďalšie, by síce zachytával variabilitu prehovorov z tohto hľadiska a poskytoval nadštandardné informácie, no vzhľadom na náročnosť ostane vymedzený pre perspektívny menší podkorpus.

Častým javom v ústnych prehovoroch sú nedokončené výpovede, ktoré vznikajú prerušením výpovede druhým účastníkom dialógu alebo pri zamyslení sa hovoriaceho, prípadne aj nahradením slovného prejavu gestom, mimikou či iným nejazykovým prejavom. Nezriedka sa stáva, že výpoveď sa končí nielen nedokončením myšlienky včlenenej do vety,

ale aj uprostred slova. Pokiaľ ide o nedokončené slovo, je praktické uviesť v zátvorke celý predpokladaný tvar.

Nevýhodou nahrávok aj pri všetkej snahe o popísanie situácie, údajoch o hovoriacich a ich vzájomného vzťahu, miesta, v ktorom sa dialóg realizuje, je neprítomnosť obrazového záznamu. V potrebných prípadoch, nie však enormne, bude potrebné zachytiť aj rušivé vplyvy, pokiaľ výrazne ovplyvňujú či narúšajú, príp. dopĺňajú komunikáciu, napr. urobiť záznam o rušivom hluku, zvukoch prístrojov, o posunkoch a pohyboch hovoriacich, ktoré zastupujú slová. Záznamy zo skupín RUCHY a VÝSLOVNOSŤ budú vkladané do prepisu v transcriberi prostredníctvom tabuľky *Vložiť udalosť* (Ctrl+e; Ctrl+d). Neartikulované zvuky sa zaznačia podľa podobnosti: viac konsonant (typu *hm, mhm, mmm*) ako MM, viac vokál (typu *ee, ehm*) ako EE.

Vo fonologickom spôsobe prepisu sa dôsledne zachytáva znelostná asimilácia (*fták, zmena, pížme, nážho*). Spodobovanie hlások sa realizuje retrográdne, a to nielen v rámci slovných tvarov na hraniciach morfematických švíkov, ale aj s presahom hraníc slovných tvarov. Vzhľadom na požiadavky softvérových nástrojov na automatickú analýzu textu a aj pre efektívnejšie vyhľadávanie slovných tvarov by bolo vhodné postupovať v tomto prípade v súlade so slovenskou pravopisnou normou, ktorá je kombináciou fonologického a etymologického spôsobu zápisu, k nim by sa mohla do zátvoriek pripisovať reálna výslovnosť.

Typické pre vokalický systém slovenčiny sú diftongy (dvojhlásky), kĺzavé vokalické zvuky zložené z dvoch rozlíšiteľných častí, ktoré ako nedeliteľný celok tvoria jadro slabiky. Vrchol sonórnosti je v druhej časti dvojhlások, ich dĺžka je porovnateľná s dlhými samohláskami. Môžu sa vyskytovať v akejkoľvek pozícii slova (*priasť, vysvedčenie, besiedka, kôň*). Treba ich odlišovať od hiátov vyskytujúcich sa v slovách cudzieho pôvodu (*akcia, diecéza, Siux*). Zdvojené spoluhlásky zapisujeme podľa toho, ako boli vyslovené (*ranný, raný, jesenný, denný*).

V slovenčine platí pravidlo o rytmickom krátení, z tohto pravidla sú však aj výnimky. Pri dôslednom zachytávaní dĺžky samohlások pri prepise budú zachytené aj všetky prípady výnimiek, pravda, s ohľadom na pôvod respondentov.

Zásady prepisu nahrávok Slovenského hovoreného korpusu sú determinované cieľom a budú sa realizovať ako 1. textová (ortografická) podoba a 2. cielená ortoepická transkripcia, pričom prepis bude k dispozícii na prehliadanie spolu so zvukom. V korpusovom manažéri Manatee s klientom Bonito bude možné vyhľadávať spôsobmi tradičnými pre písané korpusy, ale aj podľa výslovnosti, keď sa osobitne zobrazia záznamy zapísané v zátvorkách.

Ukážka prepisu podľa možnosti 3) bez zachytenia doplňujúcich informácií (šumy, ruchy...):

Na súde sa dnes začalo pojednávanie v spore medzi vládnou SNS a poslancom opozičnej SMK Miklósom Durayom, ktorý stranu Jána Slotu označil za fašistickú. Poslanec tým podľa národniarov poškodil dobré meno strany, a chce preto od neho vysúdiť desať miliónov korún. Durayov právnik však argumentoval, že ak by súd vyhovel SNS, išlo by o porušenie poslancovho práva na slobodu prejavu.

Na sú[ú]d[ď]e sa dn[ň]es[z] začalo pojedná[á]van[ň]ie[ie] v[f] spore medz[dz:]i vlá[á]dnou SNS[esenes] *(skratka)* a poslancom opozičn[n]ej SMK[esemká] *(skratka)* Mikló[ó]šom Duray[j]om, *(pauza)* ktorý[í] stranu Já[á]na Slotu označil za fašist[t]ickú[ú]. Poslan[ň]ec[dz:] t[t]ý[í]m podľ[ľ]a národn[ň]ia[ia]rov poškod[ď]il dobré[é] meno stran[n]y[i], a chce preto od[d] n[ň]eho vy[i]sú[ú]d[ď]iť[ď] d[ď]esať[ď] milió[ó]nov[u] korú[ú]n. duray[j]ov[u] prá[á]vn[ň]ik[g] však[g] argumentoval, *(pauza)* že ak by[i] sú[ú]d[d] vy[i]hovel SNS[esenes] *(skratka)*, *(pauza*) išlo by[i] o porušen[ň]ie[ie] poslancov[u]ho prá[á]va na slobodu prejavu.

LITERATÚRA

Bosák J., 1990, Skúmanie jazyka ako sociálno-komunikačného systému. In *Dynamické tendencie v jazykovej komunikácii*, red. J. Bosák, JÚĽŠ SAV, Bratislava, 75-84.

Bosák J., 1993, Skúmanie slovenčiny ako sociálno-komunikačného systému v slovanskom kontexte. *Slavica Slovaca*, 28, 171-178.

Čermák F., 2006, Mluvené korpusy. In *Korpusová lingvistika: Stav a modelové přístupy* 1, eds. F. Čermák a R. Blatná, Nakladatelství Lidové noviny/Ústav Českého národního korpusu, Praha, 53-67.

Český národní korpus. Ústav Českého národního korpusu FF UK, Praha. Dostupný z World Wide Web: <http://ucnk.ff.cuni.cz>.

Esvan F., 2006, Srovnávací rozbor mluvených korpusů (PMK a BMK): metodologické problémy a první výsledky. In *Korpusová lingvistika: Stav a modelové přístupy* 1, eds. F. Čermák a R. Blatná, Nakladatelství Lidové noviny/Ústav Českého národního korpusu, Praha, 95-117.

Horecký J., 1979, Vymedzenie štandardnej formy slovenčiny. *Slovenská reč*, 44, 221-227.

Horecký J., 1984, Na okraj štruktúrnej klasifikácie F. Kočiša. *Slovenská reč*, 49, 162-167.

Hovorená podoba spisovnej slovenčiny, I., II. Referáty a diskusné príspevky z konferencie dňa 6.-9. októbra 1965. Interný materiál Združenia slovenských jazykovedcov.

Kráľ Á., 2005, Pravidlá slovenskej výslovnosti. Systematika a ortoepický slovník. Matica slovenská, Martin.

Mesto a jeho jazyk, 2000. *Sociolinguistica Slovaca* 5, ed. S. Ondrejovič, Veda, Bratislava.

Patráš V., 1990, Hovorená podoba slovenčiny v Banskej Bystrici. [Autoreferát kandidátskej dizertačnej práce.] JÚĽŠ SAV, Bratislava.

Pořízka P., 2005, Přepis(y) textů v korpusech mluvené češtiny. In *Jazyky v kontaktu/jazyky v konfliktu a evropský jazykový prostor,* VUP, Olomouc, 235-240.

Pořízka P., 2004, K možnostem vyhledávání dat a struktuře atributů korpusových manažerů v mluvených korpusech ČNK. In AUPO *Philologica* 84, *Bohemica* IX, VUP, Olomouc, 81-92.

Slančová D., M. Sokolová, 1995, Výskum podoby hovorenej komunikácie na východnom Slovensku. In *Sociolinguistica Slovaca* 1, red. S. Ondrejovič, M. Šimková, Veda, Bratislava, 132-143.

Slovenský národný korpus. Dostupný z World Wide Web: <http://korpus.juls.savba.sk>.

VOICE. Dostupný z World Wide Web: <www.univie.ac.at/voice>.