Mária ŠIMKOVÁ

Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka

1. Slovenský národný korpus ako všeobecný jednojazyčný korpus písaných textov súčasného slovenského jazyka v elektronickej podobe sa začal tvoriť v roku 2002 v oddelení Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra SAV v Bratislave. Je jedným z najmladších nekomerčných projektov tohto druhu v strednej a východnej Európe; v skupine okolitých krajín, ktoré nemajú takýto fungujúci národný korpus, ostali po „odchode“ Slovenska už len Ukrajina a Bielorusko. Niektoré štáty (napr. Česko a Poľsko) disponujú viacerými rozsahom veľkými a významom relevantnými korpusmi. Okrem všeobecných jednojazyčných korpusov písaných textov sa v súčasnosti dostáva do popredia budovanie paralelných korpusov (dvoj- alebo viacjazyčných) a korpusov hovorených prejavov. Oproti minulosti je dnes už takmer povinnou výbavou každého korpusu aj nejaká miera lingvistických informácií – lematizácia, morfologická, syntaktická, sémantická, štýlovo-žánrová anotácia a pod. Korpusová lingvistika zaznamenala za uplynulé desaťročie búrlivý rozvoj a čoraz väčší záujem o jej výsledky prejavujú aj bádatelia v oblasti počítačového spracovania prirodzeného jazyka (natural language processing – NLP). Stalo sa už okrídlenou frázou konštatovanie, že od prvotnej otázky Načo korpus? a vysvetľovania jeho pozitívnych vlastností pri výskume jazyka sa diskusia dnes posunula najmä k problému, ako z korpusu dostať čo najviac relevantných informácií.

O vývine matematickej (a v jej rámci počítačovej) lingvistiky na Slovensku, o vzniku a plánoch Slovenského národného korpusu (ďalej SNK) sme písali na viacerých miestach (Šimková, 2003, 2004; http://korpus.juls.savba.sk/korpus/o_korpusoch; o Slovenskom národnom korpuse). V tomto príspevku sa zameriame na špecifiká budovania SNK z hľadiska (každého) potenciálneho používateľa a na súčasný stav možností využitia SNK na štúdium slovenčiny v dvoch významoch tohto slova: štúdium ako vedecké bádanie, výskum a štúdium ako učenie sa niečomu, získavanie poznatkov.

2. Budovanie SNK nadväzovalo na tradíciu predchádzajúceho korpusu textov slovenského jazyka, ktorý sa tvoril v Jazykovednom ústave Ľudovíta Štúra prakticky od roku 1993 do roku 2002 ako elektronická databáza lingvisticky nespracovaného (surového) textového materiálu predovšetkým pre potreby lexikografického kolektívu koncipujúceho nový výkladový slovník súčasného slovenského jazyka. SNK do istej miery pokračuje v primárnom zameraní na používateľa-lexikografa, no rozšíril svoj záber aj na bežného

používateľa (laickí záujemcovia o jazyk, študenti, učitelia, redaktori a iní pracovníci so slovom či textom) a na špecialistov v oblasti gramatických výskumov i v oblasti NLP.

2.1. Každá kategória používateľov má vlastné potreby, resp. očakávania, ktoré treba pri tvorbe korpusového projektu zohľadniť. Bežní používatelia najčastejšie hľadajú v korpuse slovo alebo spojenie slov, keď zápasia s nejakým pravopisným alebo štylistickým problémom (písanie cudzích slov, vlastných mien, ale aj slov s ypsilonom či mäkkým *i*, hľadanie synoným, významov nových slov a pod.). V tejto skupine sú (môžu byť) osobitnou podskupinou učitelia slovenského jazyka, ktorí môžu využívať korpus aj na prípravu najrôznejších cvičení, od pravopisných po štylistické. Na tieto potreby stačí korpus textov v surovej, lingvisticky nespracovanej podobe a takého rozsahu, aby pokryl bežnú slovnú zásobu vrátane najnovšej a bežné jazykové prostriedky (napr. predložkové väzby). Prehliadanie textu sa v tomto prípade uskutočňuje pomocou jednoduchého softvéru, ktorý je schopný vytvoriť konkordanciu, t. j. hľadané slovo alebo jazykový prostriedok umiestni uprostred vopred stanoveného kontextu, ktorý sa dá alebo nedá rozširovať.

O niečo väčšie nároky na korpus kladie používateľ-lexikograf. Kým zistí, aké výhody pre jeho prácu poskytuje elektronicky uložený a spracovateľný korpusový materiál, vydrží istý čas so „službami“ na úrovni bežného používateľa. Čím viac však narastá korpus, čo je pre lexikografa na jednej strane veľmi potrebné, tým rastú aj jeho nároky na automatizovanie niektorých procedúr. Veľký rozsah textov v korpuse totiž na druhej strane prináša nie až taký žiaduci výsledok v podobe niekoľkotisícových (niekedy aj miliónových) konkordancií bežných slov alebo jazykových prostriedkov. Tvorca slovníka teda síce uvíta čo najväčší korpus (*more data is better data*) pokrývajúci nielen bežnú slovnú zásobu toho obdobia a tej časti jazyka, ktorú lexikograficky opisuje, ale obsahujúci aj zriedkavejšie slová a jazykové prostriedky, no potrebuje k tomu už aj adekvátne softvérové vybavenie umožňujúce selekciu výsledných dát pomocou metódy náhodného výberu alebo filtrov (pozitívnych či negatívnych), usporadúvanie materiálu podľa frekvencií a distribúcií výskytov, tvorbu kolokácií a pod. Na prácu aj v tomto prípade môže stačiť korpus textov bez dodatočných lingvistických informácií, nevyhnutné sú však informácie bibliografického a niekedy aj štýlovo-žánrového charakteru.

Podstatne náročnejšou skupinou sú potenciálni používatelia korpusu spomedzi teoretických lingvistov. Boli to zväčša práve oni, ktorí zdôrazňovali otázku Načo korpus?! a ktorí aj v súčasnosti majú viacero výhrad voči využívaniu korpusov na gramatické výskumy. M. A. K. Halliday na túto tému hovorí: „Vždy sa mi zdalo, najmä odvtedy, ako som sa prvý raz pokúsil stať sa gramatikom, že gramatika je odbor s príliš veľkou teóriou a s absolútnym nedostatkom materiálu“ (Halliday, 2000, s. 137). Tí, ktorí jeho postoj neprijímajú, sú pre

korpusy predsa prínosní ako kritici a pomáhajú nastoľovať a riešiť viaceré základné otázky. Jednou je napr. otázka ako môže gramatik korpus používať. Jemu nestačí obyčajný textový materiál; na to, aby mohol využiť automatizované predspracovanie javu, ktorý skúma, potrebuje mať v texte príslušné lingvistické informácie: lexikálno-sémantické, morfologické, syntaktické a ďalšie. Uvedené informácie musí niekto do korpusu vložiť, ten niekto je omylný subjekt, ktorý vychádza z premenlivej gramatickej teórie, a ručným značkovaním nedokáže pripraviť dostatočné množstvo materiálu na získanie relevantných výsledkov. Nástroje na automatizované značkovanie slov lingvistickými charakteristikami síce dokážu spracovať veľké množstvá dát, no štatistické metódy značkovania nie sú zďaleka stopercentné. Gramatický výskum na korpusovom materiáli sa tak spočiatku zdal ako nemožný alebo aspoň neprijateľný: ak máme dostať zlý / nedostatočný materiál, tak radšej nijaký... Dodnes sú medzi odborníkmi na korpusovú problematiku zástancovia neznačkovania textov, ktoré podľa nich najlepšie vypovedajú o reálnej podobe jazyka, keď sú nenarušené subjektívnymi interpretáciami. Zároveň však v súčasnosti existuje viacero lingvisticky označkovaných korpusov (najviac pre angličtinu) a množstvo času a energie vloženej do značkovania prináša svoje výsledky napr. aj pri príprave gramatík, akou je známa Longman Grammar of Spoken and Written English (Londýn 1999).

Bádatelia v oblasti NLP majú všetky doteraz vymenované potreby (veľké množstvo textov s čo najväčším množstvom lingvistických informácií) a navyše musia riešiť problémy, ktoré pre iných nie sú problémami: ako spracovať napr. vzorce, obrázky, grafy, bibliografické odkazy a pod., ako odlíšiť bodku na konci vety od bodky za skratkou či iniciálou krstného mena, ako narábať s rôznymi typmi úvodzoviek, pomlčiek a spojovníkov, ako rozoznať a spracovať viacslovné výrazy či vlastné mená, ako pristupovať k javom a konštrukciám, ktoré sa v gramatických teóriách zatiaľ neriešili, resp. riešili len minimálne, ale reálne existujú, ako rozpoznať a eliminovať skutočné chyby a pod. (bližšie McNaught, 2000, s. 151).

Tvorca korpusu by mal brať do úvahy všetky tieto potreby a osobitnú pozornosť venovať reprezentatívnosti korpusu. Tak ako o možnostiach korpusu a potrebách používateľov existujú mnohé štúdie, podobne sa už neraz na viacerých miestach rozoberala otázka reprezentatívnosti (napr. Biber, 2000, s. 107 – 136; http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html; v obidvoch ďalšia literatúra).

2.2. Doteraz sme uvádzali, že je dobré mať čo najviac dát (ktoré súčasná počítačová technika dokáže spracovať), hoci na jednom mieste sme už naznačili, že pri vysokých frekvenciách bežných slov môže nastať problém s príliš veľkým množstvom viac-menej rovnorodého materiálu, ktorý sa nedá a ani nie je potrebné ľudsky spracovať. Menší rozsah dát, obsahujúci primerané množstvo bežných jazykových prostriedkov, zas nevyhovuje

používateľom hľadajúcim menej alebo veľmi málo frekventované javy. Korpusové štatistiky uvádzajú, že v korpuse, ktorý obsahuje 100 miliónov textových jednotiek, čo je v súčasnosti minimálna veľkosť všeobecného korpusu, sa 8 tisíc jednotiek nachádza v 95 percentách textu a zvyšných 5 percent reprezentuje 500 tisíc jednotiek. Tento problém je možné riešiť vytvorením reprezentatívneho korpusu, v ktorom by mali byť texty pokrývajúce v čo najkomplexnejšom zábere štýly a žánre, generácie a skupiny autorov, vydavateľské úzy a pod. a súčasne obsahujúce celú škálu jazykových prostriedkov.

Pri uvažovaní o reprezentatívnosti je najskôr nevyhnutné stanoviť celok, voči ktorému má byť korpus reprezentatívnym. Ak hovoríme o všeobecnom národnom korpuse, mohli by sme oprávnene predpokladať, že by mal zachytávať všetko, čo sa kedy v danom národnom jazyku vyskytlo. Nemálo dnešných korpusov tohto typu (vrátane SNK) sa však pohybuje na úrovni písaných textov, aj to prevažne synchrónnych; elektronické spracovanie spontánnych hovorených prejavov a podobne aj diachrónnych textov je ešte stále ekonomicky veľmi náročné. Všeobecný korpus zameraný na reálne proporčné zachytenie jazykového úzu by podľa analýz D. Bibera (c. d., s. 116) mal obsahovať približne 90 % rozhovorov (bežne hovorenej reči), 3 % listov a poznámok a 7 % publikovaných a nepublikovaných textov klasických štýlov a žánrov.

Aj keď sa všeobecný národný korpus zostavuje z písaných textov súčasného jazyka, pri tvorbe projektu treba jednak z časového hľadiska vymedziť súčasný jazyk a jednak z hľadiska reprezentatívnosti stanoviť, či sa texty do korpusu budú vyberať na základe princípu adekvátneho zastúpenia všetkých štýlov a žánrov (textových typov) alebo na základe adekvátneho rozloženia jazykových javov v príslušných textoch, resp. v celom korpuse (napr. v publicistických textoch môžeme očakávať veľmi nízke zastúpenie *ja*-formy, v odborných textoch technického zamerania nízku variantnosť syntaktických štruktúr a pod.). Najmä druhý princíp však predpokladá empirický výskum na rozsiahlom, jazykovo spracovanom materiáli – teda na lingvisticky označkovanom korpuse.

V prípade prijatia prvého princípu sa riešia ďalšie otázky. Majú sa texty zahrnuté do reprezentatívneho korpusu vyberať a stratifikovať podľa kritéria a) produkcie textov, b) recepcie textov alebo c) textov ako produktov (Biber, c. d., s. 111). Prvé dve kritériá by výber zúžili na malé percento autorov, resp. na dosť úzky okruh najviac čítaných textov (z publicistiky športové spravodajstvo, z beletrie zelená a červená knižnica a pod.) a boli by vhodné skôr na výstavbu špecializovaných korpusov, napr. na skúmanie úzu mladej básnickej generácie, resp. napr. na skúmanie jazyka textov, ktoré najviac ovplyvňujú deti vo veku 8 – 10 rokov (sem by sa vyberali rozprávky, učebnice, detské časopisy a pod.). V prípade tretieho kritéria prichádzajú do úvahy všetky texty vyprodukované v stanovenom období, ktoré treba

opäť nejakým spôsobom roztriediť, aby sa z nich dali vyberať reprezentanti istého textového typu. D. Biber (tamže) navrhuje použiť na hierarchizáciu úrovní výberu a na zaradenie textov do „registrov“ situačné parametre, pričom register chápe ako pojem skôr kontinuálny a nie diskrétny:

1. primárny kanál – písaný jazyk / hovorené prejavy / zapísané hovorené prejavy;
2. formát – publikovaný / nepublikovaný;
3. scéna – inštitucionálna / iná verejná / súkromná, osobná;
4. adresát – a) pluralita: ne/vymenovaný / hromadný / individuálny / autor ako adresát;
   b) prítomnosť (miesto a čas): prítomný / neprítomný;
   c) interaktívnosť: žiadna / malá / značná;
   d) zdieľaná znalosť: všeobecná / špecializovaná / osobná;
5. adresor – a) demografická variácia: pohlavie, vek, zamestnanie a i.;
   b) uznanie poplatnosti / vďačnosti: uvádzaný jednotlivec / inštitúcia;
6. faktuálnosť – faktuálno-informačná / stredná či neurčitá / imaginatívna;
7. účel – presvedčiť, pobaviť, povzniesť, informovať, poučiť, vysvetliť, rozprávať, opísať, zaznamenať, vyjadriť sám seba, vyjadriť postoj, názor alebo emócie, posilniť medziľudský vzťah ...
8. témy ...

Členenie textov na základe uvedených charakteristík sa značne líši od našej tradičnej štýlovo-žánrovej klasifikácie a najmä posledné dva parametre si vyžadujú ďalší teoretický a empirický výskum; hoci tematická klasifikácia publikovaných textov sa nachádza aj v slovenských archívnych či knižničných systémoch, hovorené prejavy sú z tohto hľadiska veľmi málo preskúmané a prepracované. Napriek tomu možnosť stratifikácie textov do registrov podľa situačných parametrov predstavuje z istého hľadiska jednoduchší postup pri stanovovaní reprezentatívnosti korpusu. Ideálny korpus by však mal reprezentovať nielen škálu registrov, ale aj škálu rôznorodosti, distribúcie jazykových prostriedkov v jednotlivých typoch textov. Z analýz D. Bibera (c. d., s. 118 – 127) vyplýva, že:

a) bežné lineárne jazykové javy majú v textoch pomerne stabilnú distribúciu a dajú sa spoľahlivo zistiť aj z relatívne krátkych textových segmentov (už aj v rozsahu 1000 slov);

b) zriedkavé jazykové javy vykazujú značnú distribučnú rôznorodosť a vyžadujú dlhšie textové vzorky;

c) javy s distribúciou pravdepodobnosti po krivke, t. j. rôzne typy javov, napr. kumulatívnosť slovných druhov, sú relatívne stabilné naprieč po sebe idúcimi textovými segmentmi, ale počet výskytov nových typov v texte postupne klesá, pričom frekvencia nových typov je sústavne vyššia v textových segmentoch naprieč textami ako vo vzorke

jedného textu. Čiže viac rôznorodosti v typoch textov zaradených do korpusu sa premietne do širšej reprezentatívnosti typov jazykových javov, pričom vzorky textov musia byť dostatočne dlhé, aby dokázali distribúciu jazykových javov spoľahlivo reprezentovať.

Rôzne korpusové projekty pristupujú k otázke výberu textov a reprezentatívnosti korpusu rôzne – od abstrahovania od tejto vlastnosti korpusu (Bank of English, Mannheimer Corpora) až po detailné prepracovanie percentuálneho zastúpenia štýlov a žánrov na základe rozsiahlych a opakovaných sociolingvistických prieskumov (Český národný korpus). Viaceré korpusové projekty si postupne vypracúvali vlastné kritériá výstavby korpusu (bližšie napr. Čermák, 2000, s. 21 – 22), ktoré sa v súčasnosti môžu objektivizovať na základe štandardizovaných odporúčaní TEI (Text Encoding Initiative) a EAGLES (Expert Advisory Group on Language Engineering Standards). Všeobecnou metodikou po doterajších skúsenostiach je však predovšetkým cyklickosť zostavovania reprezentatívneho korpusu: po teoretickom stanovení situačných parametrov determinujúcich výber textov v danej jazykovej komunite a po stanovení rozsahu dôležitých jazykových javov, ktoré sa budú v korpuse analyzovať, sa vytvorí pilotný korpus s dostatočne širokou škálou rôznorodosti, ale i hĺbkou textov a registrov. Tieto texty sa gramaticky označkujú a na pilotnom korpuse sa uskutoční empirický výskum, ktorého výsledky potvrdia alebo pozmenia použité parametre. Jednotlivé fázy tohto cyklu by mali prebiehať veľmi kontinuálne a dynamicky modifikovať celkový charakter korpusu. Takýto postup sa dodržiava napr. v Českom národnom korpuse, kde vytvoreniu prvého reprezentatívneho korpusu českého jazyka SYN2000 predchádzal predbežný prieskum, na základe ktorého bolo stanovené zastúpenie hlavných štýlov v tomto rozsahu: umelecká literatúra a literatúra faktu 15 %, noviny a časopisy 60 %, odborná literatúra 25 %. Vysoký podiel novín a časopisov bol odvodený z pôvodného prieskumu Opinion Window Prague 1996. Ukázalo sa však, že sa odvtedy jednak zmenili podmienky vydávania a následne aj čítanosti periodík (poklesli počty denníkov a časopisov a ich náklady a vzrástli ceny), jednak sa v r. 2001 robili dva nové prieskumy na základe odlišnej metodiky a pomocou rôzne kladených otázok. Výsledkom je návrh novej štruktúry Českého národného korpusu v podobe 40 % umelecká literatúra a literatúra faktu, 33 % noviny a časopisy, 27 % odborná literatúra. Administratívne texty, ktoré boli v pôvodnom rozvrhnutí začlenené medzi odborné texty, sa podľa novej metodiky do korpusu vôbec nezaraďujú vzhľadom na ich informačnú hodnotu, ktorá je v podstate rovnaká ako pri odborných textoch, používatelia s nimi pracujú opakovane ako s inštrukciami, pokynmi, predpismi a nečítajú ich systematicky ako iné druhy literatúry (Králík, 2004).

3. Pri príprave projektu Slovenského národného korpusu sme vychádzali zo skúseností existujúcich korpusových projektov, najmä českých, z potrieb potenciálnych používateľov

elektronickej databázy slovenských textov a z reálnych možností pracovného kolektívu, ktorý má minimálnu možnú zostavu (7 pracovných miest) a tvoria ho nie absolventi odborov počítačovej či korpusovej lingvistiky, pretože takéto odbory na Slovensku nie sú na žiadnej vysokej škole. Začiatkom roka 2003 sme nemali ani jeden text, ktorý by bol zmluvne dohodnutý na verejné (internetové) použitie v elektronickej databáze. V rámci prípravy koncepcie Slovenského národného korpusu na roky 2003 – 2006 (Šimková, 2003, 2004) sme vymedzili tieto základné ciele:

1. Vybudovanie všeobecného jednojazyčného korpusu písaných textov súčasného slovenského jazyka (1955–2005) a sprístupnenie jeho reprezentatívnej časti v rozsahu 200 mil. slov na internete; sprístupnenú časť korpusu lematizovať a morfologicky anotovať, vybranú vzorku anotovať aj syntakticky.
2. Sprístupnenie celého súboru elektronicky spracovaných získaných textov bez lingvistických informácií pracovníkom JÚĽŠ SAV a externým spolupracovníkom v jeho priestoroch na vedecko-výskumné, predovšetkým lexikografické účely (rozsah bude podmienený technickými možnosťami a ústretovosťou poskytovateľov textov).
3. Budovanie špecifických korpusov / databáz

– terminologická databáza (v spolupráci s Ministerstvom spravodlivosti SR a rezortnými terminologickými komisiami);

– databáza lexikografických diel (sprístupnenie lexikografickej produkcie JÚĽŠ SAV v elektronickej podobe na internete, príp. aj na CD nosičoch);

– korpus diachrónnych textov a korpus nárečových textov (podľa potrieb pracovníkov v príslušných odboroch a podľa technických možností – náročné bude najmä optické rozpoznávanie starých tlačí či rukopisov a prepisy hovorenej reči);

– paralelný/-é korpus/-y (najmä pre tzv. malé jazyky sú takéto korpusy dobrým nástrojom pre prekladateľov a tlmočníkov, ale aj dobrým prostriedkom na zviditeľňovanie a akceptáciu jazyka vo svete);

– korpus hovorených prejavov (technická a časová náročnosť prepisu bude vyžadovať samostatné finančné i personálne zabezpečenie).

4. Tvorba príslušných programových nástrojov (archivovanie textov; evidenčná databáza; konverzie a filtrácie textov; lematizátor, morfologický anotátor), využívanie a adaptácia existujúcich programových nástrojov (syntaktický parser; korpusový manažér).

Na zber dát sme stanovili princíp “čo najviac a čo najrozmanitejších textov”. O reprezentatívnej vzorke písaných textov súčasného slovenského jazyka sme uvažovali vo veľmi hrubých rysoch: jedna tretina publicistických textov, jedna tretina umeleckých textov a jedna tretina odborných a populárno-náučných textov. (Istým potvrdením správnosti tohto

prvotného základného rozvrstvenia sú aj spomínané české sociolingvistické prieskumy a návrh novej štruktúry Českého národného korpusu.) V posledných dvoch skupinách sme dôraz kládli na preklady, ktoré majú v menších národných a jazykových spoločenstvách, akým je aj slovenské spoločenstvo, osobitné postavenie, ale pri príprave doterajších lexikografických príručiek slovenského jazyka nemali takmer žiadne zastúpenie. Pre Slovenský národný korpus sme navrhovali zhruba jednotretinový podiel preložených umeleckých a odborných či populárno-náučných textov. V publicistických textoch sa preklady takisto vyskytujú, je však podstatne náročnejšie, v niektorých prípadoch takmer nemožné osobitne ich vyčleniť (preklady agentúrnych správ, pri ktorých sa neuvádza, že ide o preklad a táto informácia sa nedá automatizovane zachytiť).

3.1. Vzhľadom na akútnu potrebu materiálu pre lexikografický kolektív pripravujúci nový výkladový slovník súčasného slovenského jazyka a vzhľadom na dostupnosť a ústretovosť poskytovateľov textov sme v prvej fáze volili taktiku prijať každý text, ktorý sa podarí získať bez vynakladania prílišného úsilia (získavanie textov od kontaktovania poskytovateľa cez vysvetľovanie cieľov, obsahu a nekomerčného charakteru projektu až po uzatvorenie príslušnej zmluvy o použití diela na vedecko-výskumné účely podľa autorského zákona trvá v priemere jeden až dva mesiace). Kým sa podarilo uzavrieť východiskový počet zmlúv na zaradenie textov do korpusu, pripravovala sa metodika segmentácie (tokenizácie) slovenského textu a vonkajšej, bibliografickej a štýlovo-žánrovej anotácie (Garabík, 2004, Horák, 2004). Získané texty sa priebežne spracúvali a sprístupňovali na použitie cez internet, čo mohlo klásť isté nároky na používateľov, aby sa vedeli zorientovať, s akým rozsahom a štruktúraciou textov práve pracujú, no dôležitejšie bolo, že už mohli so slovenskými textami pracovať (počet zmluvne registrovaných používateľov zo Slovenska, ale i zo zahraničia sa pohybuje okolo 100). Po prvej verzii sprístupnenej v auguste 2003 a obsahujúcej 26 miliónov slov nasledovala v decembri 2003 druhá verzia, ktorá obsahovala 166 miliónov slov, v súčasnosti sa sprístupňuje tretia verzia s novou tokenizáciou a prepracovanou štýlovo-žánrovou anotáciou (podrobnejšie Horák, 2004), ktorá bude mať vyše 215 miliónov slov, z toho vyše 204 miliónov verejne sprístupnených na internete. Spočiatku sa v korpuse vyhľadávalo prostredníctvom jednoduchého www rozhrania (základné vyhľadávanie bez podpory regulárnych výrazov a bez zobrazovania vonkajšej anotácie) a na základe zmluvne dohodnutých podmienok aj prostredníctvom korpusového manažéra Manatee s klientom Bonito, ktorý bol vyvinutý na Fakulte informatiky Masarykovej univerzity v Brne (pozri obrázok s konkordanciou tvarov slova *(naj)onakvejší* v závere príspevku). Nová verzia SNK sa dá prehľadávať pomocou nového korpusového manažéra Korman, ktorý sa vyvíja v oddelení Slovenského národného korpusu JÚĽŠ SAV a plánuje sa perspektívne aplikovať aj

na paralelné korpusy. Štruktúra nových dát SNK predstavuje takmer 199 miliónov textových jednotiek z publicistických textov, vyše 7 miliónov umeleckých textov a vyše 3 milióny odborných a populárno-náučných textov. Disproporcia v prospech publicistických textov je výrazná, no súčasné zastúpenie nepublicistických textov už umožňuje uvažovať o tvorbe vyváženého korpusu, ktorý by mal v rámci základnej tretinovej štrukturácie po doplnení odborných textov zhruba 30 miliónov tokenov. Na takomto korpuse, ktorého texty sú aj hlbšie žánrovo anotované, by sa dali robiť empirické výskumy vedúce k tvorbe reprezentatívneho korpusu písaných textov súčasného slovenského jazyka. Vyvažovanie korpusových textov je nevyhnutné aj pre potreby morfosyntaktického výskumu slovenčiny na korpusovom materiáli (grantový projekt Vega v spolupráci s FF PU Prešov), v rámci ktorého sa zároveň bude skúmať aj distribúcia jazykových javov v jednotlivých typoch textov, čo následne ovplyvní reprezentatívny výber textov z lingvistického hľadiska. Zistená frekvencia textových jednotiek po lematizácii súčasnej štruktúry textov v SNK však už teraz vykazuje štandardné rozloženie najfrekventovanejších predložiek, spojok, zámen a častíc, ako je známe aj z predchádzajúcich výskumov (pozri zoznam 50 najfrekventovanejších lem v závere príspevku).

3.2. V doterajšej práci na Slovenskom národnom korpuse neostáva bokom ani podiel lingvistickej zložky, tá však vzhľadom na svoj charakter postupuje podstatne pomalšie a prvé relevantné výsledky budeme môcť očakávať v priebehu roka 2005. Od začiatku r. 2003 sa vypracúvali pravidlá morfologickej anotácie (Forróová – Horák, 2004, Garabík – Gianitsová – Horák – Šimková, 2004), v súčasnosti prebieha ručné anotovanie, pri ktorom sa využíva aj práca študentov bratislavských a prešovských filologických katedier. Zároveň sa korigujú výsledky automatizovanej morfologickej anotácie uskutočňovanej pomocou nástrojov vyvinutých na počítačovo- a korpusovolingvistických pracoviskách v Čechách. Ručne anotovaný a dezambiguovaný korpus by mal mať rozsah min. 1 milión slovných tvarov, aby bol dostatočným trénovacím korpusom pre vlastný slovenský anotačný nástroj. Pri výbere textov na ručnú anotáciu sme takisto zohľadňovali potrebu rôznorodosti – v prvej fáze sa anotuje jeden umelecký preklad a vybrané texty z internetového časopisu InZine, nasledovať budú texty z odbornej a populárno-náučnej literatúry.

4. Slovenský národný korpus poskytuje v súčasnom stave základný výskumný materiál pre všetky kategórie používateľov, resp. záujemcov o slovenský jazyk. Nenahrádza však pravopisné či gramatické príručky; predstavuje len východisko na ich tvorbu, ale východisko s dobrou dostupnosťou v rámci práce prostredníctvom internetu a s podstatne väčšími možnosťami v rámci automatizovaného spracovania veľkého množstva reálnych textov. Po dopracovaní naznačených úloh bude postupne napĺňať kritériá kladené na reprezentatívny,

resp. aspoň vyvážený korpus národného jazyka všeobecného typu. Ako každý takýto korpus však bude vždy odrážať iba reálny výskyt jazykových prostriedkov nachádzajúcich sa v textoch, ktoré sú v korpuse spracované, ale neposkytne informáciu o neexistujúcich jazykových prostriedkoch, pretože korpus síce môže obsahovať veľké množstvo materiálu, no zrejme neobsiahne celý jazykový systém.

Literatúra

BIBER, D.: *Reprezentativnost v projektu korpusu*. In: *Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4*. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 107 – 136.

ČERMÁK, F.: *Jazykový korpus: prostředek a zdroj poznání*. In: *Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4*. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 15 – 37.

FORRÓOVÁ, M. – HORÁK, A.: *Morfologická anotácia korpusu*. In: *Slovenčina na začiatku 21. storočia*. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 174 – 183.

GARABÍK, R.: *Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia*. In: *Slovenčina na začiatku 21. storočia*. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 164 – 173; aktuálna verzia http://korpus.juls.savba.sk.

GARABÍK, R. – GIANITSOVÁ, L. – HORÁK, A. – ŠIMKOVÁ, M.: *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu*. In: http://korpus.juls.savba.sk, 2004.

HORÁK, A.: *Style-Genre Annotation of the Slovak National Corpus.* In: *Corpora and Corpus Linguistics*. Ed. M. Šimková. Bratislava: Veda 2004 (v tlači).

HALLIDAY, M. A. K.: *Jazyk jako systém a jazyk jako instance: korpus jako teoretický konstrukt.* In: *Studie z korpusové lingvistiky*. *Acta Universitatis Carolinae. Philologica 3 – 4*. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 137 – 150.

KRÁLÍK, J.: *Aktualizace rozvržení zdrojů Českého národního korpusu s ohledem na revizi vyváženosti jeho struktury.* Slovo a slovesnost, 65, 2004, , č. 2, s. 133 – 141.

McNAUGHT, J.: *Potřeby uživatelů textových korpusů při počítačovém zpracování přirozeného jazyka.* In: *Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4*. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 151 – 165.

ŠIMKOVÁ, M.: *Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus*. In: *Počítačová podpora prekladu. Zborník prednášok (Budmerice 22. – 23. máj 2003).* Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.

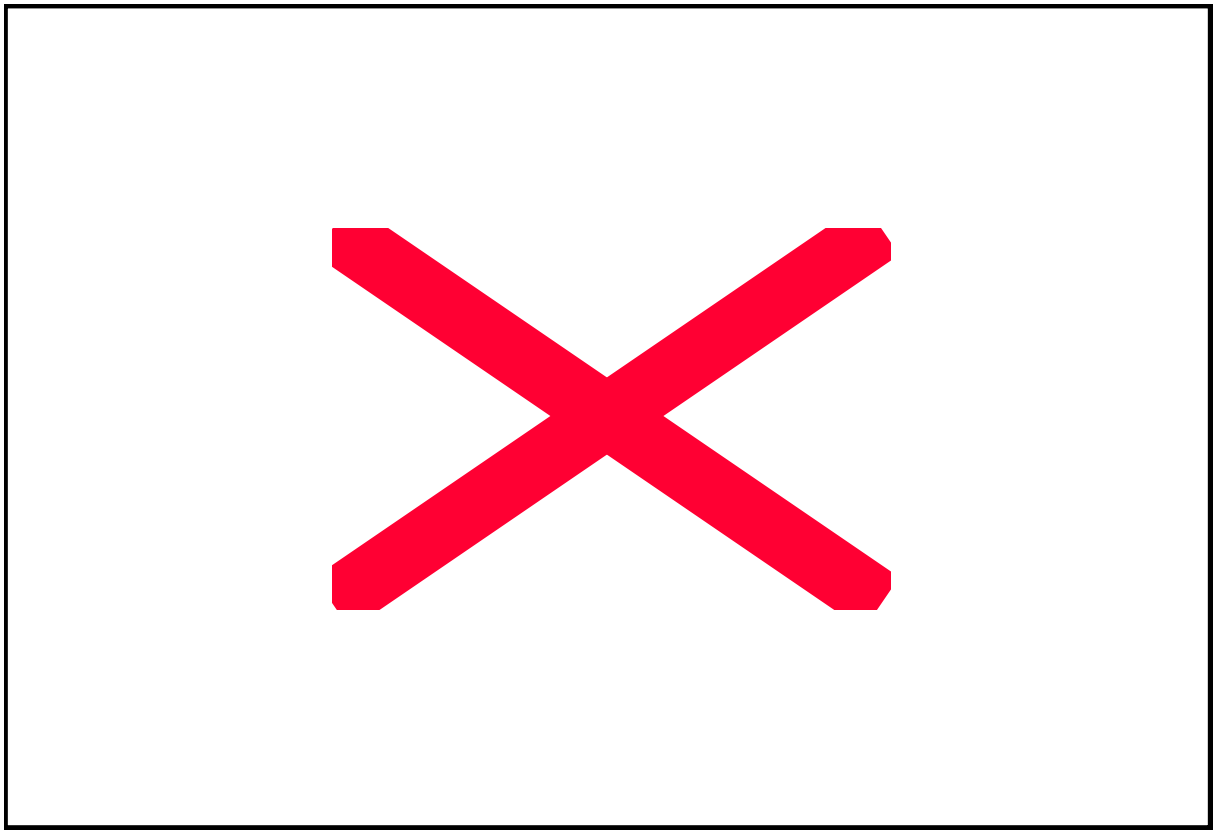
ŠIMKOVÁ, M.: *Slovenský národný korpus – východiská a plány*. In*: Slovenčina na začiatku 21. storočia.* Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.

http://korpus.juls.savba.sk

http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Tento príspevok vznikol v rámci úlohy štátneho programu Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu a v rámci grantového projektu MŠ SR VEGA 1/3149/04 Morfosyntaktický výskum v rámci Slovenského národného korpusu.

Korpusový manažér Manatee s klientom Bonito: ukážka konkordancie tvarov slova *(naj)onakvejší* s morfologickou anotáciou



Zoznam 50 najfrekventovanejších textových jednotiek v Slovenskom národnom korpuse

.	12593940
,	12024917
byť	4173315
v	4116967
a	3733556
sa	3430393
na	2876376
:	1702654
ten	1665043
)	1511721
(	1500963
"	1442677
ktorý	1392608
s	1345996
z	1284548

že	1269601
on	1119011
o	1111194
mať	963093
aj	948124
do	879967
ako	755018
tento	683605
rok	642159
1	594908
ja	587736
za	568525
po	558161
jeho	553985
by	547309
svoj	498688
pre	490304
00	485852
už	467173
ale	461627
k	456006
2	442745
čo	421128
môcť	417130
však	379155
od	356145
„	354036
0	349993
3	348247
keď	333421
?	321304
podľa	314977
môj	310901
30	309404
všetko	309335